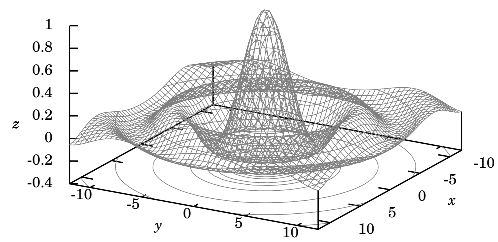


Vector Calculus

Michael Corral

Vector Calculus

Michael Corral

Schoolcraft College

About the author:
Michael Corral is an Adjunct Faculty member of the Department of Mathematics at Schoolcraft College. He received a B.A. in Mathematics from the University of California at Berkeley, and received an M.A. in Mathematics and an M.S. in Industrial & Operations Engineering from the University of Michigan.

This text was typeset in LaTeX 2_ε with the KOMA-Script bundle, using the GNU Emacs text editor on a Fedora Linux system. The graphics were created using MetaPost, PGF, and Gnuplot.

Preface

This book covers calculus in two and three variables. It is suitable for a one-semester course, normally known as “Vector Calculus”, “Multivariable Calculus”, or simply “Calculus III”. The prerequisites are the standard courses in single-variable calculus (a.k.a. Calculus I and II).

I have tried to be somewhat rigorous about proving results. But while it is important for students to see full-blown proofs - since that is how mathematics works - too much rigor and emphasis on proofs can impede the flow of learning for the vast majority of the audience at this level. If I were to rate the level of rigor in the book on a scale of 1 to 10, with 1 being completely informal and 10 being completely rigorous, I would rate it as a 5.

There are 420 exercises throughout the text, which in my experience are more than enough for a semester course in this subject. There are exercises at the end of each section, divided into three categories: A, B and C. The A exercises are mostly of a routine computational nature, the B exercises are slightly more involved, and the C exercises usually require some effort or insight to solve. A crude way of describing A, B and C would be “Easy”, “Moderate” and “Challenging”, respectively. However, many of the B exercises are easy and not all the C exercises are difficult.

There are a few exercises that require the student to write his or her own computer program to solve some numerical approximation problems (e.g. the Monte Carlo method for approximating multiple integrals, in Section 3.4). The code samples in the text are in the Java programming language, hopefully with enough comments so that the reader can figure out what is being done even without knowing Java. Those exercises do not mandate the use of Java, so students are free to implement the solutions using the language of their choice. While it would have been simple to use a scripting language like Python, and perhaps even easier with a functional programming language (such as Haskell or Scheme), Java was chosen due to its ubiquity, relatively clear syntax, and easy availability for multiple platforms.

Answers and hints to most odd-numbered and some even-numbered exercises are provided in Appendix A. Appendix B contains a proof of the right-hand rule for the cross product, which seems to have virtually disappeared from calculus texts over the last few decades. Appendix C contains a brief tutorial on Gnuplot for graphing functions of two variables.

This book is released under the GNU Free Documentation License (GFDL), which allows others to not only copy and distribute the book but also to modify it. For more details, see the included copy of the GFDL. So that there is no ambiguity on this matter, anyone can make as many copies of this book as desired and distribute it as desired, without needing my permission. The PDF version will always be freely available to the public at no cost (go to `http://www.mecmath.net`). Feel free to contact me at `mcorral@schoolcraft.edu` for

any questions on this or any other matter involving the book (e.g. comments, suggestions, corrections, etc). I welcome your input.

Finally, I would like to thank my students in Math 240 for being the guinea pigs for the initial draft of this book, and for finding the numerous errors and typos it contained.

January 2008 MICHAEL CORRAL

Contents

1 Vectors in Euclidean Space

1.1 Introduction

In single-variable calculus, the functions that one encounters are functions of a variable (usually x or t) that varies over some subset of the real number line (which we denote by $\mathbb{R}$). For such a function, say, $y = f(x)$, the **graph** of the function f consists of the points $(x, y) = (x, f(x))$. These points lie in the **Euclidean plane**, which, in the **Cartesian** or **rectangular** coordinate system, consists of all ordered pairs of real numbers (a, b). We use the word "Euclidean" to denote a system in which all the usual rules of Euclidean geometry hold. We denote the Euclidean plane by $\mathbb{R}^2$; the "2" represents the number of *dimensions* of the plane. The Euclidean plane has two perpendicular **coordinate axes**: the x-axis and the y-axis.

In vector (or multivariable) calculus, we will deal with functions of two or three variables (usually x, y or x, y, z, respectively). The graph of a function of two variables, say, $z = f(x, y)$, lies in **Euclidean space**, which in the Cartesian coordinate system consists of all ordered triples of real numbers (a, b, c). Since Euclidean space is 3-dimensional, we denote it by $\mathbb{R}^3$. The graph of f consists of the points $(x, y, z) = (x, y, f(x, y))$. The 3-dimensional coordinate system of Euclidean space can be represented on a flat surface, such as this page or a blackboard, only by giving the illusion of three dimensions, in the manner shown in Figure 1.1.1. Euclidean space has three mutually perpendicular coordinate axes (x, y and z), and three mutually perpendicular coordinate planes: the xy-plane, yz-plane and xz-plane (see Figure 1.1.2).

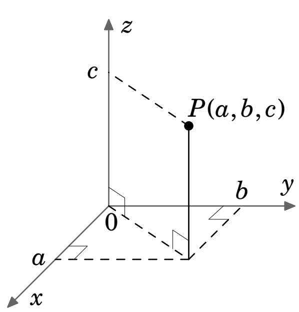


Figure 1.1.1

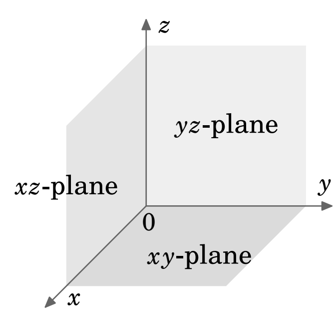


Figure 1.1.2

The coordinate system shown in Figure 1.1.1 is known as a **right-handed coordinate system**, because it is possible, using the right hand, to point the index finger in the positive direction of the x-axis, the middle finger in the positive direction of the y-axis, and the thumb in the positive direction of the z-axis, as in Figure 1.1.3.

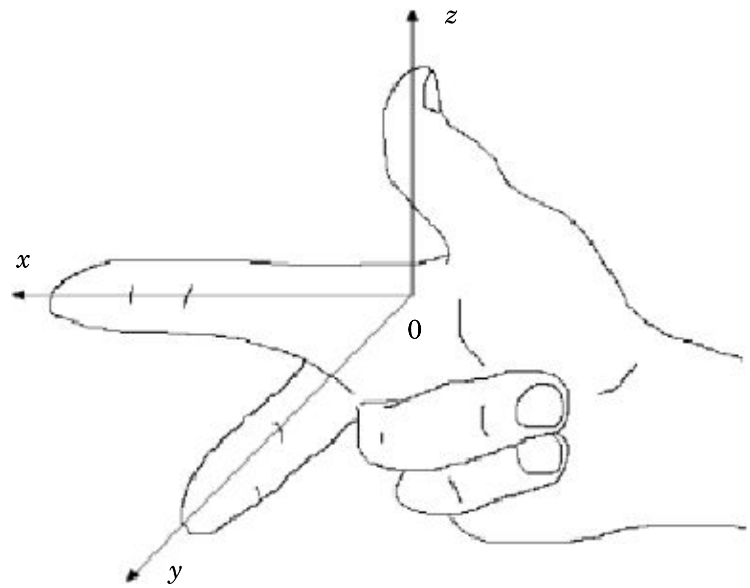


Figure 1.1.3 Right-handed coordinate system

An equivalent way of defining a right-handed system is if you can point your thumb upwards in the positive z-axis direction while using the remaining four fingers to rotate the x-axis towards the y-axis. Doing the same thing with the left hand is what defines a **left-handed coordinate system**. Notice that switching the x- and y-axes in a right-handed system results in a left-handed system, and that rotating either type of system does not change its “handedness”. Throughout the book we will use a right-handed system.

For functions of three variables, the graphs exist in 4-dimensional space (i.e. $\mathbb{R}^4$), which we can not see in our 3-dimensional space, let alone simulate in 2-dimensional space. So we can only think of 4-dimensional space abstractly. For an entertaining discussion of this subject, see the book by ABBOTT.[1]

So far, we have discussed the *position* of an object in 2-dimensional or 3-dimensional space. But what about something such as the velocity of the object, or its acceleration? Or the gravitational force acting on the object? These phenomena all seem to involve motion and *direction* in some way. This is where the idea of a *vector* comes in.

[1]One thing you will learn is why a 4-dimensional creature would be able to reach inside an egg and remove the yolk without cracking the shell!

You have already dealt with velocity and acceleration in single-variable calculus. For example, for motion along a straight line, if $y = f(t)$ gives the displacement of an object after time t, then $dy/dt = f'(t)$ is the velocity of the object at time t. The derivative $f'(t)$ is just a number, which is positive if the object is moving in an agreed-upon "positive" direction, and negative if it moves in the opposite of that direction. So you can think of that number, which was called the velocity of the object, as having two components: a *magnitude*, indicated by a nonnegative number, preceded by a *direction*, indicated by a plus or minus symbol (representing motion in the positive direction or the negative direction, respectively), i.e. $f'(t) = \pm a$ for some number $a \ge 0$. Then a is the magnitude of the velocity (normally called the *speed* of the object), and the $\pm$ represents the direction of the velocity (though the + is usually omitted for the positive direction).

For motion along a straight line, i.e. in a 1-dimensional space, the velocities are also contained in that 1-dimensional space, since they are just numbers. For general motion along a curve in 2- or 3-dimensional space, however, velocity will need to be represented by a multidimensional object which should have both a magnitude and a direction. A geometric object which has those features is an arrow, which in elementary geometry is called a "directed line segment". This is the motivation for how we will define a vector.

Definition 1.1. A (nonzero) **vector** is a directed line segment drawn from a point P (called its **initial point**) to a point Q (called its **terminal point**), with P and Q being distinct points. The vector is denoted by $\overrightarrow{PQ}$. Its **magnitude** is the length of the line segment, denoted by $\|\overrightarrow{PQ}\|$, and its **direction** is the same as that of the directed line segment. The **zero vector** is just a point, and it is denoted by $\mathbf{0}$.

To indicate the direction of a vector, we draw an arrow from its initial point to its terminal point. We will often denote a vector by a single bold-faced letter (e.g. $\mathbf{v}$) and use the terms "magnitude" and "length" interchangeably. Note that our definition could apply to systems with any number of dimensions (see Figure 1.1.4 (a)-(c)).

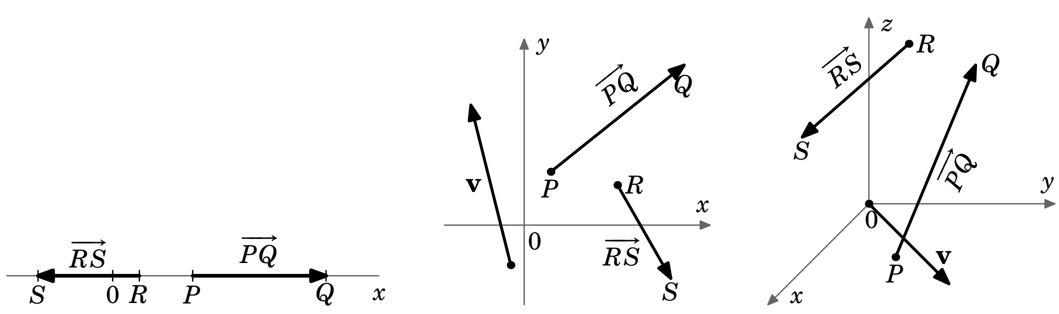


Figure 1.1.4 Vectors in different dimensions

A few things need to be noted about the zero vector. Our motivation for what a vector is included the notions of magnitude and direction. What is the magnitude of the zero vector? We define it to be zero, i.e. $\|\mathbf{0}\| = 0$. This agrees with the definition of the zero vector as just a point, which has zero length. What about the direction of the zero vector? A single point really has no well-defined direction. Notice that we were careful to only define the direction of a *nonzero* vector, which is well-defined since the initial and terminal points are distinct. Not everyone agrees on the direction of the zero vector. Some contend that the zero vector has *arbitrary* direction (i.e. can take any direction), some say that it has *indeterminate* direction (i.e. the direction can not be determined), while others say that it has *no* direction. Our definition of the zero vector, however, does not require it to have a direction, and we will leave it at that.[2]

Now that we know what a vector is, we need a way of determining when two vectors are equal. This leads us to the following definition.

Definition 1.2. Two nonzero vectors are **equal** if they have the same magnitude and the same direction. Any vector with zero magnitude is equal to the zero vector.

By this definition, vectors with the same magnitude and direction but with different initial points would be equal. For example, in Figure 1.1.5 the vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ all have the same magnitude $\sqrt{5}$ (by the Pythagorean Theorem). And we see that $\mathbf{u}$ and $\mathbf{w}$ are parallel, since they lie on lines having the same slope $\frac{1}{2}$, and they point in the same direction. So $\mathbf{u} = \mathbf{w}$, even though they have different initial points. We also see that $\mathbf{v}$ is parallel to $\mathbf{u}$ but points in the opposite direction. So $\mathbf{u} \neq \mathbf{v}$.

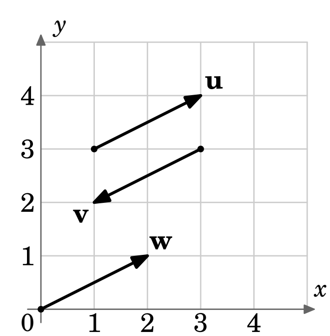


Figure 1.1.5

So we can see that there are an infinite number of vectors for a given magnitude and direction, those vectors all being equal and differing only by their initial and terminal points. Is there a single vector which we can choose to represent all those equal vectors? The answer is yes, and is suggested by the vector $\mathbf{w}$ in Figure 1.1.5.

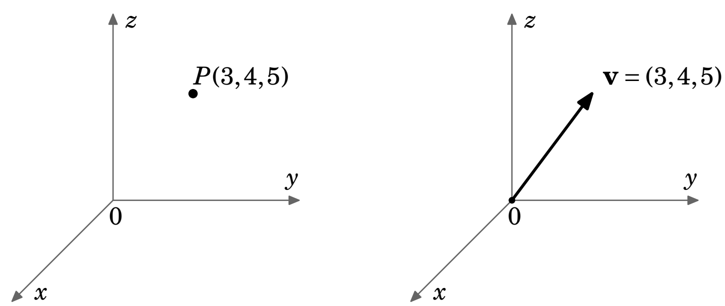
[2] In the subject of linear algebra there is a more abstract way of defining a vector where the concept of "direction" is not really used. See ANTON and RORRES.

Unless otherwise indicated, when speaking of "the vector" with a given magnitude and direction, we will mean the one whose initial point is at the origin of the coordinate system.

Thinking of vectors as starting from the origin provides a way of dealing with vectors in a standard way, since every coordinate system has an origin. But there will be times when it is convenient to consider a different initial point for a vector (for example, when adding vectors, which we will do in the next section).

Another advantage of using the origin as the initial point is that it provides an easy correspondence between a vector and its terminal point.

Example 1.1. Let $\mathbf{v}$ be the vector in $\mathbb{R}^3$ whose initial point is at the origin and whose terminal point is $(3,4,5)$. Though the *point* $(3,4,5)$ and the vector $\mathbf{v}$ are different objects, it is convenient to write $\mathbf{v} = (3,4,5)$. When doing this, it is understood that the initial point of $\mathbf{v}$ is at the origin $(0,0,0)$ and the terminal point is $(3,4,5)$.



(a) The point (3,4,5) **(b)** The vector (3,4,5)

Figure 1.1.6 Correspondence between points and vectors

Unless otherwise stated, when we refer to vectors as $\mathbf{v} = (a,b)$ in $\mathbb{R}^2$ or $\mathbf{v} = (a,b,c)$ in $\mathbb{R}^3$, we mean vectors in Cartesian coordinates starting at the origin. Also, we will write the zero vector $\mathbf{0}$ in $\mathbb{R}^2$ and $\mathbb{R}^3$ as $(0,0)$ and $(0,0,0)$, respectively.

The point-vector correspondence provides an easy way to check if two vectors are equal, without having to determine their magnitude and direction. Similar to seeing if two points are the same, you are now seeing if the terminal points of vectors starting at the origin are the same. For each vector, find the (unique!) vector it equals whose initial point is the origin. Then compare the coordinates of the terminal points of these "new" vectors: if those coordinates are the same, then the original vectors are equal. To get the "new" vectors starting at the origin, you *translate* each vector to start at the origin by subtracting the coordinates of the original initial point from the original terminal point. The resulting point will be the terminal point of the "new" vector whose initial point is the origin. Do this for each original vector then compare.

Example 1.2. Consider the vectors $\overrightarrow{PQ}$ and $\overrightarrow{RS}$ in $\mathbb{R}^3$, where $P = (2,1,5), Q = (3,5,7), R = (1,-3,-2)$ and $S = (2,1,0)$. Does $\overrightarrow{PQ} = \overrightarrow{RS}$?
Solution: The vector $\overrightarrow{PQ}$ is equal to the vector $\mathbf{v}$ with initial point $(0,0,0)$ and terminal point $Q - P = (3,5,7) - (2,1,5) = (3-2, 5-1, 7-5) = (1,4,2)$.
Similarly, $\overrightarrow{RS}$ is equal to the vector $\mathbf{w}$ with initial point $(0,0,0)$ and terminal point $S - R = (2,1,0) - (1,-3,-2) = (2-1, 1-(-3), 0-(-2)) = (1,4,2)$.
So $\overrightarrow{PQ} = \mathbf{v} = (1,4,2)$ and $\overrightarrow{RS} = \mathbf{w} = (1,4,2)$.
$\therefore \overrightarrow{PQ} = \overrightarrow{RS}$

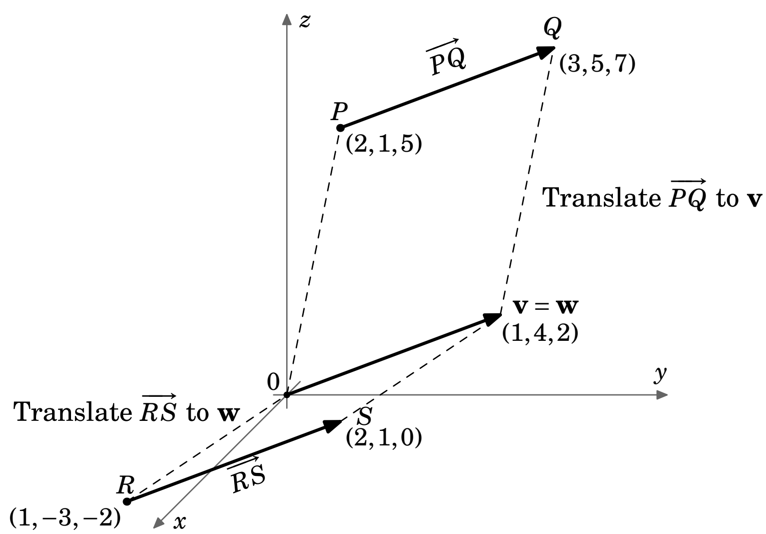


Figure 1.1.7

Recall the distance formula for points in the Euclidean plane:

For points $P = (x_1, y_1)$, $Q = (x_2, y_2)$ in $\mathbb{R}^2$, the distance d between P and Q is:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1.1}$$

By this formula, we have the following result:

For a vector $\overrightarrow{PQ}$ in $\mathbb{R}^2$ with initial point $P = (x_1, y_1)$ and terminal point $Q = (x_2, y_2)$, the magnitude of $\overrightarrow{PQ}$ is:

$$\|\overrightarrow{PQ}\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1.2}$$

Finding the magnitude of a vector $\mathbf{v} = (a, b)$ in $\mathbb{R}^2$ is a special case of formula (1.2) with $P = (0,0)$ and $Q = (a, b)$:

For a vector $\mathbf{v} = (a, b)$ in $\mathbb{R}^2$, the magnitude of $\mathbf{v}$ is:

$$\|\mathbf{v}\| = \sqrt{a^2 + b^2} \tag{1.3}$$

To calculate the magnitude of vectors in $\mathbb{R}^3$, we need a distance formula for points in Euclidean space (we will postpone the proof until the next section):

Theorem 1.1. The distance d between points $P = (x_1, y_1, z_1)$ and $Q = (x_2, y_2, z_2)$ in $\mathbb{R}^3$ is:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{1.4}$$

The proof will use the following result:

Theorem 1.2. For a vector $\mathbf{v} = (a, b, c)$ in $\mathbb{R}^3$, the magnitude of $\mathbf{v}$ is:

$$\|\mathbf{v}\| = \sqrt{a^2 + b^2 + c^2} \tag{1.5}$$

Proof: There are four cases to consider:
Case 1: $a = b = c = 0$. Then $\mathbf{v} = \mathbf{0}$, so $\|\mathbf{v}\| = 0 = \sqrt{0^2 + 0^2 + 0^2} = \sqrt{a^2 + b^2 + c^2}$.

Case 2: exactly two of a, b, c *are* 0. Without loss of generality, we assume that $a = b = 0$ and $c \neq 0$ (the other two possibilities are handled in a similar manner). Then $\mathbf{v} = (0, 0, c)$, which is a vector of length $|c|$ along the z-axis. So $\|\mathbf{v}\| = |c| = \sqrt{c^2} = \sqrt{0^2 + 0^2 + c^2} = \sqrt{a^2 + b^2 + c^2}$.

Case 3: exactly one of a, b, c *is* 0. Without loss of generality, we assume that $a = 0$, $b \neq 0$ and $c \neq 0$ (the other two possibilities are handled in a similar manner). Then $\mathbf{v} = (0, b, c)$, which is a vector in the yz-plane, so by the Pythagorean Theorem we have $\|\mathbf{v}\| = \sqrt{b^2 + c^2} = \sqrt{0^2 + b^2 + c^2} = \sqrt{a^2 + b^2 + c^2}$.

Case 4: none of a, b, c *are* 0. Without loss of generality, we can assume that a, b, c are all positive (the other seven possibilities are handled in a similar manner). Consider the points $P = (0, 0, 0)$, $Q = (a, b, c)$, $R = (a, b, 0)$, and $S = (a, 0, 0)$, as shown in Figure 1.1.8. Applying the Pythagorean Theorem to the right triangle $\triangle PSR$ gives $|PR|^2 = a^2 + b^2$. A second application of the Pythagorean Theorem, this time to the right triangle $\triangle PQR$, gives $\|\mathbf{v}\| = |PQ| = \sqrt{|PR|^2 + |QR|^2} = \sqrt{a^2 + b^2 + c^2}$.
This proves the theorem. **QED**

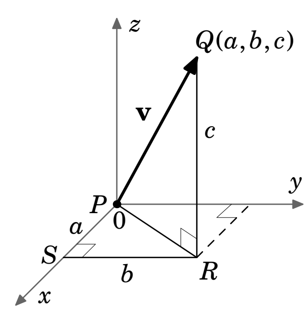


Figure 1.1.8

Example 1.3. Calculate the following:

(a) The magnitude of the vector $\overrightarrow{PQ}$ in $\mathbb{R}^2$ with $P = (-1,2)$ and $Q = (5,5)$.
Solution: By formula (1.2), $\|\overrightarrow{PQ}\| = \sqrt{(5-(-1))^2+(5-2)^2} = \sqrt{36+9} = \sqrt{45} = 3\sqrt{5}$.

(b) The magnitude of the vector $\mathbf{v} = (8,3)$ in $\mathbb{R}^2$.
Solution: By formula (1.3), $\|\mathbf{v}\| = \sqrt{8^2+3^2} = \sqrt{73}$.

(c) The distance between the points $P = (2,-1,4)$ and $Q = (4,2,-3)$ in $\mathbb{R}^2$.
Solution: By formula (1.4), the distance $d = \sqrt{(4-2)^2+(2-(-1))^2+(-3-4)^2} = \sqrt{4+9+49} = \sqrt{62}$.

(d) The magnitude of the vector $\mathbf{v} = (5,8,-2)$ in $\mathbb{R}^3$.
Solution: By formula (1.5), $\|\mathbf{v}\| = \sqrt{5^2+8^2+(-2)^2} = \sqrt{25+64+4} = \sqrt{93}$.

Exercises

A

1. Calculate the magnitudes of the following vectors:
(a) $\mathbf{v} = (2,-1)$ (b) $\mathbf{v} = (2,-1,0)$ (c) $\mathbf{v} = (3,2,-2)$ (d) $\mathbf{v} = (0,0,1)$ (e) $\mathbf{v} = (6,4,-4)$

2. For the points $P = (1,-1,1)$, $Q = (2,-2,2)$, $R = (2,0,1)$, $S = (3,-1,2)$, does $\overrightarrow{PQ} = \overrightarrow{RS}$?

3. For the points $P = (0,0,0)$, $Q = (1,3,2)$, $R = (1,0,1)$, $S = (2,3,4)$, does $\overrightarrow{PQ} = \overrightarrow{RS}$?

B

4. Let $\mathbf{v} = (1,0,0)$ and $\mathbf{w} = (a,0,0)$ be vectors in $\mathbb{R}^3$. Show that $\|\mathbf{w}\| = |a|\,\|\mathbf{v}\|$.

5. Let $\mathbf{v} = (a,b,c)$ and $\mathbf{w} = (3a,3b,3c)$ be vectors in $\mathbb{R}^3$. Show that $\|\mathbf{w}\| = 3\,\|\mathbf{v}\|$.

C

6. Though we will see a simple proof of Theorem 1.1 in the next section, it is possible to prove it using methods similar to those in the proof of Theorem 1.2. Prove the special case of Theorem 1.1 where the points $P = (x_1, y_1, z_1)$ and $Q = (x_2, y_2, z_2)$ satisfy the following conditions:
$x_2 > x_1 > 0$, $y_2 > y_1 > 0$, and $z_2 > z_1 > 0$.
(*Hint: Think of Case 4 in the proof of Theorem 1.2, and consider Figure 1.1.9.*)

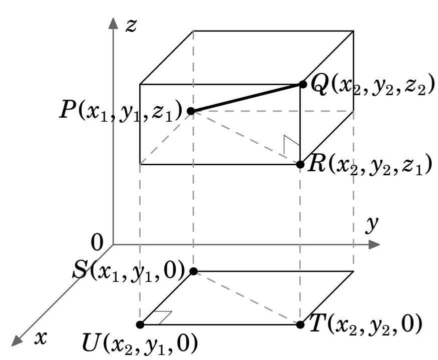


Figure 1.1.9

1.2 Vector Algebra

Now that we know what vectors are, we can start to perform some of the usual algebraic operations on them (e.g. addition, subtraction). Before doing that, we will introduce the notion of a *scalar*.

Definition 1.3. A **scalar** is a quantity that can be represented by a single number.

For our purposes, scalars will always be real numbers.[3] Examples of scalar quantities are mass, electric charge, and speed (not velocity).[4] We can now define *scalar multiplication* of a vector.

Definition 1.4. For a scalar k and a nonzero vector $\mathbf{v}$, the **scalar multiple** of $\mathbf{v}$ by k, denoted by $k\mathbf{v}$, is the vector whose magnitude is $|k|\,\|\mathbf{v}\|$, points in the same direction as $\mathbf{v}$ if $k > 0$, points in the opposite direction as $\mathbf{v}$ if $k < 0$, and is the zero vector $\mathbf{0}$ if $k = 0$. For the zero vector $\mathbf{0}$, we define $k\mathbf{0} = \mathbf{0}$ for any scalar k.

Two vectors $\mathbf{v}$ and $\mathbf{w}$ are **parallel** (denoted by $\mathbf{v} \parallel \mathbf{w}$) if one is a scalar multiple of the other. You can think of scalar multiplication of a vector as stretching or shrinking the vector, and as flipping the vector in the opposite direction if the scalar is a negative number (see Figure 1.2.1).

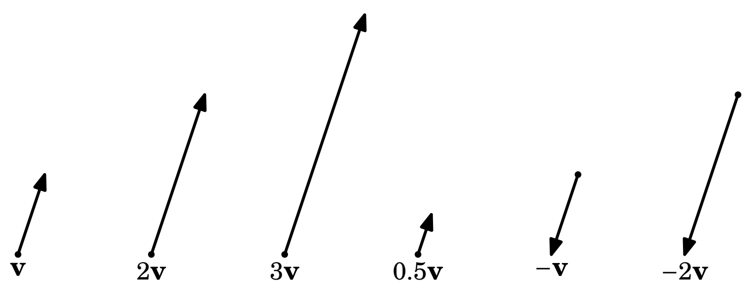


Figure 1.2.1

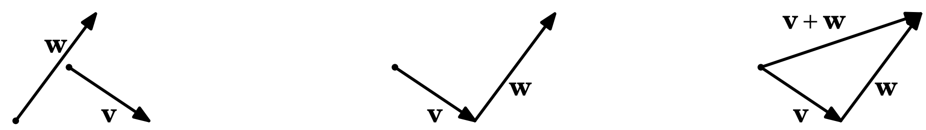
Recall that **translating** a nonzero vector means that the initial point of the vector is changed but the magnitude and direction are preserved. We are now ready to define the *sum* of two vectors.

Definition 1.5. The **sum** of vectors $\mathbf{v}$ and $\mathbf{w}$, denoted by $\mathbf{v}+\mathbf{w}$, is obtained by translating $\mathbf{w}$ so that its initial point is at the terminal point of $\mathbf{v}$; the initial point of $\mathbf{v}+\mathbf{w}$ is the initial point of $\mathbf{v}$, and its terminal point is the new terminal point of $\mathbf{w}$.

[3] The term *scalar* was invented by 19$^{\text{th}}$ century Irish mathematician, physicist and astronomer William Rowan Hamilton, to convey the sense of something that could be represented by a point on a scale or graduated ruler. The word vector comes from Latin, where it means "carrier".

[4] An alternate definition of scalars and vectors, used in physics, is that under certain types of coordinate transformations (e.g. rotations), a quantity that is not affected is a scalar, while a quantity that is affected (in a certain way) is a vector. See MARION for details.

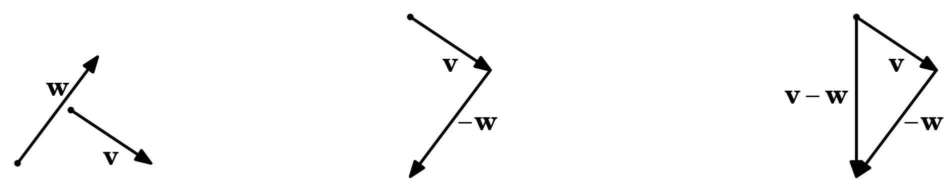
Intuitively, adding **w** to **v** means tacking on **w** to the end of **v** (see Figure 1.2.2).



(a) Vectors **v** and **w** **(b)** Translate **w** to the end of **v** **(c)** The sum **v** + **w**

Figure 1.2.2 Adding vectors **v** and **w**

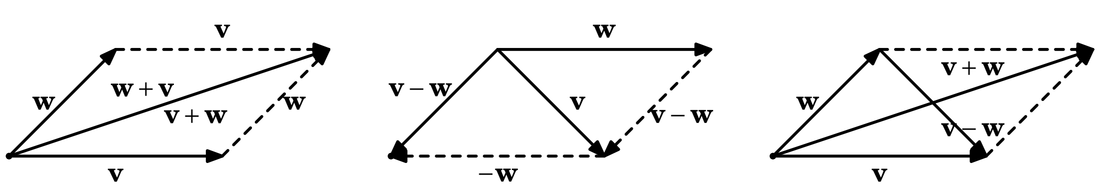
Notice that our definition is valid for the zero vector (which is just a point, and hence can be translated), and so we see that $\mathbf{v}+\mathbf{0}=\mathbf{v}=\mathbf{0}+\mathbf{v}$ for any vector **v**. In particular, $\mathbf{0}+\mathbf{0}=\mathbf{0}$. Also, it is easy to see that $\mathbf{v}+(-\mathbf{v})=\mathbf{0}$, as we would expect. In general, since the scalar multiple $-\mathbf{v}=-1\,\mathbf{v}$ is a well-defined vector, we can define **vector subtraction** as follows: $\mathbf{v}-\mathbf{w}=\mathbf{v}+(-\mathbf{w})$. See Figure 1.2.3.



(a) Vectors **v** and **w** **(b)** Translate −**w** to the end of **v** **(c)** The difference **v** − **w**

Figure 1.2.3 Subtracting vectors **v** and **w**

Figure 1.2.4 shows the use of "geometric proofs" of various laws of vector algebra, that is, it uses laws from elementary geometry to prove statements about vectors. For example, (a) shows that $\mathbf{v}+\mathbf{w}=\mathbf{w}+\mathbf{v}$ for any vectors **v**, **w**. And (c) shows how you can think of $\mathbf{v}-\mathbf{w}$ as the vector that is tacked on to the end of **w** to add up to **v**.



(a) Add vectors **(b)** Subtract vectors **(c)** Combined add/subtract

Figure 1.2.4 "Geometric" vector algebra

Notice that we have temporarily abandoned the practice of starting vectors at the origin. In fact, we have not even mentioned coordinates in this section so far. Since we will deal mostly with Cartesian coordinates in this book, the following two theorems are useful for performing vector algebra on vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$ starting at the origin.

Theorem 1.3. Let $\mathbf{v} = (v_1, v_2)$, $\mathbf{w} = (w_1, w_2)$ be vectors in $\mathbb{R}^2$, and let k be a scalar. Then
(a) $k\mathbf{v} = (kv_1, kv_2)$
(b) $\mathbf{v} + \mathbf{w} = (v_1 + w_1, v_2 + w_2)$

Proof: (a) Without loss of generality, we assume that $v_1, v_2 > 0$ (the other possibilities are handled in a similar manner). If $k = 0$ then $k\mathbf{v} = 0\mathbf{v} = \mathbf{0} = (0,0) = (0v_1, 0v_2) = (kv_1, kv_2)$, which is what we needed to show. If $k \neq 0$, then (kv_1, kv_2) lies on a line with slope $\frac{kv_2}{kv_1} = \frac{v_2}{v_1}$, which is the same as the slope of the line on which $\mathbf{v}$ (and hence $k\mathbf{v}$) lies, and (kv_1, kv_2) points in the same direction on that line as $k\mathbf{v}$. Also, by formula (1.3) the magnitude of (kv_1, kv_2) is $\sqrt{(kv_1)^2 + (kv_2)^2} = \sqrt{k^2 v_1^2 + k^2 v_2^2} = \sqrt{k^2(v_1^2 + v_2^2)} = |k|\sqrt{v_1^2 + v_2^2} = |k|\,\|\mathbf{v}\|$. So $k\mathbf{v}$ and (kv_1, kv_2) have the same magnitude and direction. This proves (a).

(b) Without loss of generality, we assume that $v_1, v_2, w_1, w_2 > 0$ (the other possibilities are handled in a similar manner). From Figure 1.2.5, we see that when translating $\mathbf{w}$ to start at the end of $\mathbf{v}$, the new terminal point of $\mathbf{w}$ is $(v_1 + w_1, v_2 + w_2)$, so by the definition of $\mathbf{v} + \mathbf{w}$ this must be the terminal point of $\mathbf{v} + \mathbf{w}$. This proves (b). **QED**

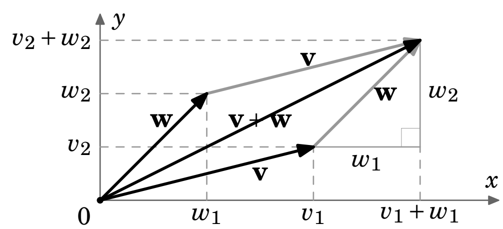


Figure 1.2.5

Theorem 1.4. Let $\mathbf{v} = (v_1, v_2, v_3)$, $\mathbf{w} = (w_1, w_2, w_3)$ be vectors in $\mathbb{R}^3$, let k be a scalar. Then
(a) $k\mathbf{v} = (kv_1, kv_2, kv_3)$
(b) $\mathbf{v} + \mathbf{w} = (v_1 + w_1, v_2 + w_2, v_3 + w_3)$

The following theorem summarizes the basic laws of vector algebra.

Theorem 1.5. For any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, and scalars k, l, we have

(a) $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$	Commutative Law
(b) $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$	Associative Law
(c) $\mathbf{v} + \mathbf{0} = \mathbf{v} = \mathbf{0} + \mathbf{v}$	Additive Identity
(d) $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$	Additive Inverse
(e) $k(l\mathbf{v}) = (kl)\mathbf{v}$	Associative Law
(f) $k(\mathbf{v} + \mathbf{w}) = k\mathbf{v} + k\mathbf{w}$	Distributive Law
(g) $(k + l)\mathbf{v} = k\mathbf{v} + l\mathbf{v}$	Distributive Law

Proof: (a) We already presented a geometric proof of this in Figure 1.2.4(a).

(b) To illustrate the difference between analytic proofs and geometric proofs in vector algebra, we will present both types here. For the analytic proof, we will use vectors in $\mathbb{R}^3$ (the proof for $\mathbb{R}^2$ is similar).

Let $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$, $\mathbf{w} = (w_1, w_2, w_3)$ be vectors in $\mathbb{R}^3$. Then

$$\begin{aligned}
\mathbf{u} + (\mathbf{v} + \mathbf{w}) &= (u_1, u_2, u_3) + ((v_1, v_2, v_3) + (w_1, w_2, w_3)) \\
&= (u_1, u_2, u_3) + (v_1 + w_1, v_2 + w_2, v_3 + w_3) && \text{by Theorem 1.4(b)} \\
&= (u_1 + (v_1 + w_1), u_2 + (v_2 + w_2), u_3 + (v_3 + w_3)) && \text{by Theorem 1.4(b)} \\
&= ((u_1 + v_1) + w_1, (u_2 + v_2) + w_2, (u_3 + v_3) + w_3) && \text{by properties of real numbers} \\
&= (u_1 + v_1, u_2 + v_2, u_3 + v_3) + (w_1, w_2, w_3) && \text{by Theorem 1.4(b)} \\
&= (\mathbf{u} + \mathbf{v}) + \mathbf{w}
\end{aligned}$$

This completes the analytic proof of (b). Figure 1.2.6 provides the geometric proof.

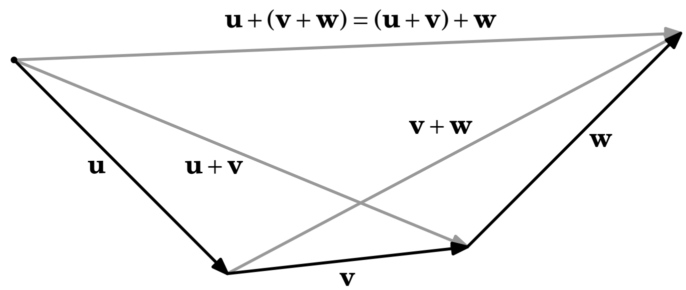


Figure 1.2.6 Associative Law for vector addition

(c) We already discussed this on p.10.

(d) We already discussed this on p.10.

(e) We will prove this for a vector $\mathbf{v} = (v_1, v_2, v_3)$ in $\mathbb{R}^3$ (the proof for $\mathbb{R}^2$ is similar):

$$\begin{aligned}
k(l\mathbf{v}) &= k(lv_1, lv_2, lv_3) && \text{by Theorem 1.4(a)} \\
&= (klv_1, klv_2, klv_3) && \text{by Theorem 1.4(a)} \\
&= (kl)(v_1, v_2, v_3) && \text{by Theorem 1.4(a)} \\
&= (kl)\mathbf{v}
\end{aligned}$$

(f) and (g): Left as exercises for the reader. **QED**

A **unit vector** is a vector with magnitude 1. Notice that for any nonzero vector $\mathbf{v}$, the vector $\frac{\mathbf{v}}{\|\mathbf{v}\|}$ is a unit vector which points in the same direction as $\mathbf{v}$, since $\frac{1}{\|\mathbf{v}\|} > 0$ and $\left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\| = \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|} = 1$. Dividing a nonzero vector $\mathbf{v}$ by $\|\mathbf{v}\|$ is often called *normalizing* $\mathbf{v}$.

There are specific unit vectors which we will often use, called the **basis vectors**: $\mathbf{i} = (1,0,0)$, $\mathbf{j} = (0,1,0)$, and $\mathbf{k} = (0,0,1)$ in $\mathbb{R}^3$; $\mathbf{i} = (1,0)$ and $\mathbf{j} = (0,1)$ in $\mathbb{R}^2$.
These are useful for several reasons: they are mutually perpendicular, since they lie on distinct coordinate axes; they are all unit vectors: $\|\mathbf{i}\| = \|\mathbf{j}\| = \|\mathbf{k}\| = 1$; every vector can be written as a unique scalar combination of the basis vectors: $\mathbf{v} = (a,b) = a\,\mathbf{i} + b\,\mathbf{j}$ in $\mathbb{R}^2$, $\mathbf{v} = (a,b,c) = a\,\mathbf{i} + b\,\mathbf{j} + c\,\mathbf{k}$ in $\mathbb{R}^3$. See Figure 1.2.7.

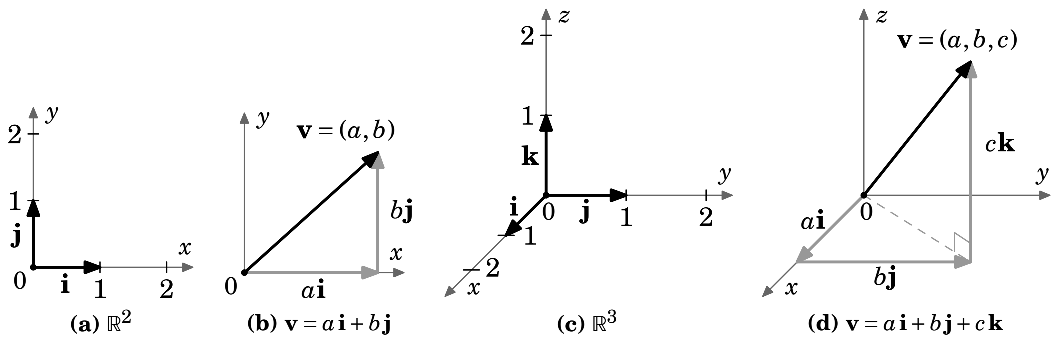


Figure 1.2.7 Basis vectors in different dimensions

When a vector $\mathbf{v} = (a, b, c)$ is written as $\mathbf{v} = a\,\mathbf{i} + b\,\mathbf{j} + c\,\mathbf{k}$, we say that $\mathbf{v}$ is in *component form*, and that a, b, and c are the ***i, j, and k components***, respectively, of $\mathbf{v}$. We have:

$$\mathbf{v} = v_1\,\mathbf{i} + v_2\,\mathbf{j} + v_3\,\mathbf{k}, k \text{ a scalar} \Longrightarrow k\mathbf{v} = kv_1\,\mathbf{i} + kv_2\,\mathbf{j} + kv_3\,\mathbf{k}$$

$$\mathbf{v} = v_1\,\mathbf{i} + v_2\,\mathbf{j} + v_3\,\mathbf{k}, \mathbf{w} = w_1\,\mathbf{i} + w_2\,\mathbf{j} + w_3\,\mathbf{k} \Longrightarrow \mathbf{v} + \mathbf{w} = (v_1 + w_1)\mathbf{i} + (v_2 + w_2)\mathbf{j} + (v_3 + w_3)\mathbf{k}$$

$$\mathbf{v} = v_1\,\mathbf{i} + v_2\,\mathbf{j} + v_3\,\mathbf{k} \Longrightarrow \|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + v_3^2}$$

Example 1.4. Let $\mathbf{v} = (2, 1, -1)$ and $\mathbf{w} = (3, -4, 2)$ in $\mathbb{R}^3$.

(a) Find $\mathbf{v} - \mathbf{w}$.
Solution: $\mathbf{v} - \mathbf{w} = (2 - 3, 1 - (-4), -1 - 2) = (-1, 5, -3)$

(b) Find $3\mathbf{v} + 2\mathbf{w}$.
Solution: $3\mathbf{v} + 2\mathbf{w} = (6, 3, -3) + (6, -8, 4) = (12, -5, 1)$

(c) Write $\mathbf{v}$ and $\mathbf{w}$ in component form.
Solution: $\mathbf{v} = 2\,\mathbf{i} + \mathbf{j} - \mathbf{k}$, $\mathbf{w} = 3\,\mathbf{i} - 4\mathbf{j} + 2\,\mathbf{k}$

(d) Find the vector $\mathbf{u}$ such that $\mathbf{u} + \mathbf{v} = \mathbf{w}$.
Solution: By Theorem 1.5, $\mathbf{u} = \mathbf{w} - \mathbf{v} = -(\mathbf{v} - \mathbf{w}) = -(-1, 5, -3) = (1, -5, 3)$, by part(a).

(e) Find the vector $\mathbf{u}$ such that $\mathbf{u} + \mathbf{v} + \mathbf{w} = \mathbf{0}$.
Solution: By Theorem 1.5, $\mathbf{u} = -\mathbf{w} - \mathbf{v} = -(3, -4, 2) - (2, 1, -1) = (-5, 3, -1)$.

(f) Find the vector $\mathbf{u}$ such that $2\mathbf{u} + \mathbf{i} - 2\mathbf{j} = \mathbf{k}$.
Solution: $2\mathbf{u} = -\mathbf{i} + 2\mathbf{j} + \mathbf{k} \Longrightarrow \mathbf{u} = -\frac{1}{2}\,\mathbf{i} + \mathbf{j} + \frac{1}{2}\,\mathbf{k}$

(g) Find the unit vector $\frac{\mathbf{v}}{\|\mathbf{v}\|}$.
Solution: $\frac{\mathbf{v}}{\|\mathbf{v}\|} = \frac{1}{\sqrt{2^2+1^2+(-1)^2}}(2, 1, -1) = \left(\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}\right)$

We can now easily prove Theorem 1.1 from the previous section. The distance d between two points $P = (x_1, y_1, z_1)$ and $Q = (x_2, y_2, z_2)$ in $\mathbb{R}^3$ is the same as the length of the vector $\mathbf{w} - \mathbf{v}$, where the vectors $\mathbf{v}$ and $\mathbf{w}$ are defined as $\mathbf{v} = (x_1, y_1, z_1)$ and $\mathbf{w} = (x_2, y_2, z_2)$ (see Figure 1.2.8). So since $\mathbf{w} - \mathbf{v} = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$, then $d = \|\mathbf{w} - \mathbf{v}\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$ by Theorem 1.2.

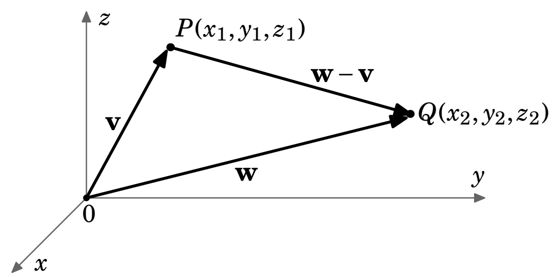


Figure 1.2.8 Proof of Theorem 1.2: $d = \|\mathbf{w} - \mathbf{v}\|$

Exercises

A

1. Let $\mathbf{v} = (-1, 5, -2)$ and $\mathbf{w} = (3, 1, 1)$.

(a) Find $\mathbf{v} - \mathbf{w}$. (b) Find $\mathbf{v} + \mathbf{w}$. (c) Find $\frac{\mathbf{v}}{\|\mathbf{v}\|}$. (d) Find $\left\|\frac{1}{2}(\mathbf{v} - \mathbf{w})\right\|$.

(e) Find $\left\|\frac{1}{2}(\mathbf{v} + \mathbf{w})\right\|$. (f) Find $-2\mathbf{v} + 4\mathbf{w}$. (g) Find $\mathbf{v} - 2\mathbf{w}$.

(h) Find the vector $\mathbf{u}$ such that $\mathbf{u} + \mathbf{v} + \mathbf{w} = \mathbf{i}$.

(i) Find the vector $\mathbf{u}$ such that $\mathbf{u} + \mathbf{v} + \mathbf{w} = 2\mathbf{j} + \mathbf{k}$.

(j) Is there a scalar m such that $m(\mathbf{v} + 2\mathbf{w}) = \mathbf{k}$? If so, find it.

2. For the vectors $\mathbf{v}$ and $\mathbf{w}$ from Exercise 1, is $\|\mathbf{v} - \mathbf{w}\| = \|\mathbf{v}\| - \|\mathbf{w}\|$? If not, which quantity is larger?

3. For the vectors $\mathbf{v}$ and $\mathbf{w}$ from Exercise 1, is $\|\mathbf{v} + \mathbf{w}\| = \|\mathbf{v}\| + \|\mathbf{w}\|$? If not, which quantity is larger?

B

4. Prove Theorem 1.5(f) for $\mathbb{R}^3$.

5. Prove Theorem 1.5(g) for $\mathbb{R}^3$.

C

6. We know that every vector in $\mathbb{R}^3$ can be written as a scalar combination of the vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$. Can every vector in $\mathbb{R}^3$ be written as a scalar combination of just $\mathbf{i}$ and $\mathbf{j}$, i.e. for any vector $\mathbf{v}$ in $\mathbb{R}^3$, are there scalars m, n such that $\mathbf{v} = m\,\mathbf{i} + n\,\mathbf{j}$? Justify your answer.

1.3 Dot Product

You may have noticed that while we did define multiplication of a vector by a scalar in the previous section on vector algebra, we did not define multiplication of a vector by a vector. We will now see one type of multiplication of vectors, called the *dot product*.

Definition 1.6. Let $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$ be vectors in $\mathbb{R}^3$.
The **dot product** of $\mathbf{v}$ and $\mathbf{w}$, denoted by $\mathbf{v} \cdot \mathbf{w}$, is given by:

$$\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 + v_3 w_3 \tag{1.6}$$

Similarly, for vectors $\mathbf{v} = (v_1, v_2)$ and $\mathbf{w} = (w_1, w_2)$ in $\mathbb{R}^2$, the dot product is:

$$\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 \tag{1.7}$$

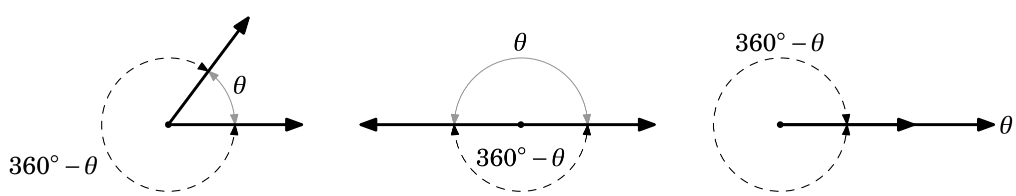
Notice that the dot product of two vectors is a scalar, not a vector. So the associative law that holds for multiplication of numbers and for addition of vectors (see Theorem 1.5(b),(e)), does *not* hold for the dot product of vectors. Why? Because for vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, the dot product $\mathbf{u} \cdot \mathbf{v}$ is a scalar, and so $(\mathbf{u} \cdot \mathbf{v}) \cdot \mathbf{w}$ is not defined since the left side of that dot product (the part in parentheses) is a scalar and not a vector.

For vectors $\mathbf{v} = v_1\,\mathbf{i} + v_2\,\mathbf{j} + v_3\,\mathbf{k}$ and $\mathbf{w} = w_1\,\mathbf{i} + w_2\,\mathbf{j} + w_3\,\mathbf{k}$ in component form, the dot product is still $\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 + v_3 w_3$.

Also notice that we defined the dot product in an analytic way, i.e. by referencing vector coordinates. There is a geometric way of defining the dot product, which we will now develop as a consequence of the analytic definition.

Definition 1.7. The **angle** between two nonzero vectors with the same initial point is the smallest angle between them.

We do not define the angle between the zero vector and any other vector. Any two nonzero vectors with the same initial point have two angles between them: θ and $360° - \theta$. We will always choose the smallest nonnegative angle θ between them, so that $0° \le \theta \le 180°$. See Figure 1.3.1.



(a) $0° < \theta < 180°$ **(b)** $\theta = 180°$ **(c)** $\theta = 0°$

Figure 1.3.1 Angle between vectors

We can now take a more geometric view of the dot product by establishing a relationship between the dot product of two vectors and the angle between them.

Theorem 1.6. Let $\mathbf{v}$, $\mathbf{w}$ be nonzero vectors, and let θ be the angle between them. Then

$$\cos\theta = \frac{\mathbf{v}\cdot\mathbf{w}}{\|\mathbf{v}\|\,\|\mathbf{w}\|} \tag{1.8}$$

Proof: We will prove the theorem for vectors in $\mathbb{R}^3$ (the proof for $\mathbb{R}^2$ is similar). Let $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$. By the Law of Cosines (see Figure 1.3.2), we have

$$\|\mathbf{v}-\mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2\,\|\mathbf{v}\|\,\|\mathbf{w}\|\cos\theta \tag{1.9}$$

(note that equation (1.9) holds even for the "degenerate" cases $\theta = 0°$ and $180°$).

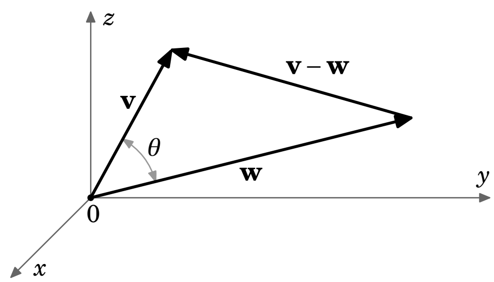


Figure 1.3.2

Since $\mathbf{v}-\mathbf{w} = (v_1 - w_1, v_2 - w_2, v_3 - w_3)$, expanding $\|\mathbf{v}-\mathbf{w}\|^2$ in equation (1.9) gives

$$\begin{aligned}
\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2\,\|\mathbf{v}\|\,\|\mathbf{w}\|\cos\theta &= (v_1 - w_1)^2 + (v_2 - w_2)^2 + (v_3 - w_3)^2 \\
&= (v_1^2 - 2v_1w_1 + w_1^2) + (v_2^2 - 2v_2w_2 + w_2^2) + (v_3^2 - 2v_3w_3 + w_3^2) \\
&= (v_1^2 + v_2^2 + v_3^2) + (w_1^2 + w_2^2 + w_3^2) - 2(v_1w_1 + v_2w_2 + v_3w_3) \\
&= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2(\mathbf{v}\cdot\mathbf{w})\text{ , so} \\
-2\,\|\mathbf{v}\|\,\|\mathbf{w}\|\cos\theta &= -2(\mathbf{v}\cdot\mathbf{w})\text{ , so since } \mathbf{v} \neq \mathbf{0} \text{ and } \mathbf{w} \neq \mathbf{0} \text{ then} \\
\cos\theta &= \frac{\mathbf{v}\cdot\mathbf{w}}{\|\mathbf{v}\|\,\|\mathbf{w}\|}\text{ , since } \|\mathbf{v}\| > 0 \text{ and } \|\mathbf{w}\| > 0.
\end{aligned}$$

QED

Example 1.5. Find the angle θ between the vectors $\mathbf{v} = (2, 1, -1)$ and $\mathbf{w} = (3, -4, 1)$.

Solution: Since $\mathbf{v}\cdot\mathbf{w} = (2)(3) + (1)(-4) + (-1)(1) = 1$, $\|\mathbf{v}\| = \sqrt{6}$, and $\|\mathbf{w}\| = \sqrt{26}$, then

$$\cos\theta = \frac{\mathbf{v}\cdot\mathbf{w}}{\|\mathbf{v}\|\,\|\mathbf{w}\|} = \frac{1}{\sqrt{6}\sqrt{26}} = \frac{1}{2\sqrt{39}} \approx 0.08 \implies \theta = 85.41°$$

Two nonzero vectors are **perpendicular** if the angle between them is $90°$. Since $\cos 90° = 0$, we have the following important corollary to Theorem 1.6:

Corollary 1.7. Two nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ are perpendicular if and only if $\mathbf{v}\cdot\mathbf{w} = 0$.

We will write $\mathbf{v} \perp \mathbf{w}$ to indicate that $\mathbf{v}$ and $\mathbf{w}$ are perpendicular.

Since $\cos\theta > 0$ for $0° \le \theta < 90°$ and $\cos\theta < 0$ for $90° < \theta \le 180°$, we also have:

Corollary 1.8. If θ is the angle between nonzero vectors $\mathbf{v}$ and $\mathbf{w}$, then

$$\mathbf{v}\cdot\mathbf{w} \text{ is } \begin{cases} >0 & \text{for } 0° \le \theta < 90° \\ 0 & \text{for } \theta = 90° \\ <0 & \text{for } 90° < \theta \le 180° \end{cases}$$

By Corollary 1.8, the dot product can be thought of as a way of telling if the angle between two vectors is acute, obtuse, or a right angle, depending on whether the dot product is positive, negative, or zero, respectively. See Figure 1.3.3.

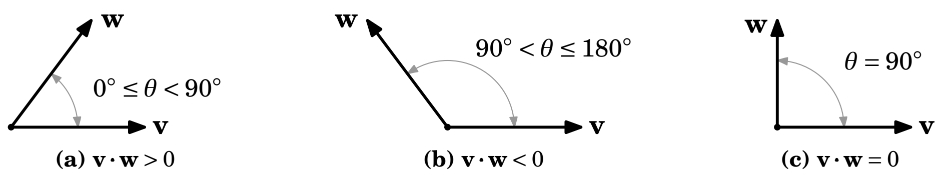


Figure 1.3.3 Sign of the dot product & angle between vectors

Example 1.6. Are the vectors $\mathbf{v} = (-1,5,-2)$ and $\mathbf{w} = (3,1,1)$ perpendicular?

Solution: Yes, $\mathbf{v} \perp \mathbf{w}$ since $\mathbf{v}\cdot\mathbf{w} = (-1)(3)+(5)(1)+(-2)(1) = 0$.

The following theorem summarizes the basic properties of the dot product.

Theorem 1.9. For any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, and scalar k, we have

(a) $\mathbf{v}\cdot\mathbf{w} = \mathbf{w}\cdot\mathbf{v}$ Commutative Law

(b) $(k\mathbf{v})\cdot\mathbf{w} = \mathbf{v}\cdot(k\mathbf{w}) = k(\mathbf{v}\cdot\mathbf{w})$ Associative Law

(c) $\mathbf{v}\cdot\mathbf{0} = 0 = \mathbf{0}\cdot\mathbf{v}$

(d) $\mathbf{u}\cdot(\mathbf{v}+\mathbf{w}) = \mathbf{u}\cdot\mathbf{v}+\mathbf{u}\cdot\mathbf{w}$ Distributive Law

(e) $(\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = \mathbf{u}\cdot\mathbf{w}+\mathbf{v}\cdot\mathbf{w}$ Distributive Law

(f) $|\mathbf{v}\cdot\mathbf{w}| \le \|\mathbf{v}\|\,\|\mathbf{w}\|$ Cauchy-Schwarz Inequality[5]

Proof: The proofs of parts (a)-(e) are straightforward applications of the definition of the dot product, and are left to the reader as exercises. We will prove part (f).

(f) If either $\mathbf{v} = \mathbf{0}$ or $\mathbf{w} = \mathbf{0}$, then $\mathbf{v}\cdot\mathbf{w} = 0$ by part (c), and so the inequality holds trivially. So assume that $\mathbf{v}$ and $\mathbf{w}$ are nonzero vectors. Then by Theorem 1.6,

$$\begin{aligned} \mathbf{v}\cdot\mathbf{w} &= \cos\theta\,\|\mathbf{v}\|\,\|\mathbf{w}\|\text{ , so} \\ |\mathbf{v}\cdot\mathbf{w}| &= |\cos\theta|\,\|\mathbf{v}\|\,\|\mathbf{w}\|\text{ , so} \\ |\mathbf{v}\cdot\mathbf{w}| &\le \|\mathbf{v}\|\,\|\mathbf{w}\| \text{ since } |\cos\theta| \le 1. \end{aligned}$$

QED

[5] Also known as the Cauchy-Schwarz-Buniakovski Inequality.

Using Theorem 1.9, we see that if $\mathbf{u}\cdot\mathbf{v} = 0$ and $\mathbf{u}\cdot\mathbf{w} = 0$, then $\mathbf{u}\cdot(k\mathbf{v}+l\mathbf{w}) = k(\mathbf{u}\cdot\mathbf{v})+l(\mathbf{u}\cdot\mathbf{w}) = k(0)+l(0) = 0$ for all scalars k, l. Thus, we have the following fact:

If $\mathbf{u} \perp \mathbf{v}$ and $\mathbf{u} \perp \mathbf{w}$, then $\mathbf{u} \perp (k\mathbf{v}+l\mathbf{w})$ for all scalars k, l.

For vectors $\mathbf{v}$ and $\mathbf{w}$, the collection of all scalar combinations $k\mathbf{v}+l\mathbf{w}$ is called the **span** of $\mathbf{v}$ and $\mathbf{w}$. If nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ are parallel, then their span is a line; if they are not parallel, then their span is a plane. So what we showed above is that a vector which is perpendicular to two other vectors is also perpendicular to their span.

The dot product can be used to derive properties of the magnitudes of vectors, the most important of which is the *Triangle Inequality*, as given in the following theorem:

Theorem 1.10. For any vectors $\mathbf{v}$, $\mathbf{w}$, we have
(a) $\|\mathbf{v}\|^2 = \mathbf{v}\cdot\mathbf{v}$
(b) $\|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\|+\|\mathbf{w}\|$ Triangle Inequality
(c) $\|\mathbf{v}-\mathbf{w}\| \ge \|\mathbf{v}\|-\|\mathbf{w}\|$

Proof: (a) Left as an exercise for the reader.

(b) By part (a) and Theorem 1.9, we have

$$\begin{aligned}\|\mathbf{v}+\mathbf{w}\|^2 &= (\mathbf{v}+\mathbf{w})\cdot(\mathbf{v}+\mathbf{w}) = \mathbf{v}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{w}+\mathbf{w}\cdot\mathbf{v}+\mathbf{w}\cdot\mathbf{w}\\ &= \|\mathbf{v}\|^2+2(\mathbf{v}\cdot\mathbf{w})+\|\mathbf{w}\|^2 \text{ , so since } a \le |a| \text{ for any real number } a, \text{ we have}\\ &\le \|\mathbf{v}\|^2+2\,|\mathbf{v}\cdot\mathbf{w}|+\|\mathbf{w}\|^2 \text{ , so by Theorem 1.9(f) we have}\\ &\le \|\mathbf{v}\|^2+2\,\|\mathbf{v}\|\,\|\mathbf{w}\|+\|\mathbf{w}\|^2 = (\|\mathbf{v}\|+\|\mathbf{w}\|)^2 \text{ and so}\end{aligned}$$

$\|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\|+\|\mathbf{w}\|$ after taking square roots of both sides, which proves (b).

(c) Since $\mathbf{v} = \mathbf{w}+(\mathbf{v}-\mathbf{w})$, then $\|\mathbf{v}\| = \|\mathbf{w}+(\mathbf{v}-\mathbf{w})\| \le \|\mathbf{w}\|+\|\mathbf{v}-\mathbf{w}\|$ by the Triangle Inequality, so subtracting $\|\mathbf{w}\|$ from both sides gives $\|\mathbf{v}\|-\|\mathbf{w}\| \le \|\mathbf{v}-\mathbf{w}\|$. **QED**

The Triangle Inequality gets its name from the fact that in any triangle, no one side is longer than the sum of the lengths of the other two sides (see Figure 1.3.4). Another way of saying this is with the familiar statement "the shortest distance between two points is a straight line."

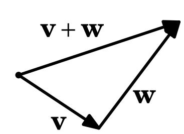


Figure 1.3.4

Exercises

A

1. Let $\mathbf{v} = (5,1,-2)$ and $\mathbf{w} = (4,-4,3)$. Calculate $\mathbf{v}\cdot\mathbf{w}$.

2. Let $\mathbf{v} = -3\,\mathbf{i}-2\,\mathbf{j}-\mathbf{k}$ and $\mathbf{w} = 6\,\mathbf{i}+4\,\mathbf{j}+2\,\mathbf{k}$. Calculate $\mathbf{v}\cdot\mathbf{w}$.

For Exercises 3-8, find the angle θ between the vectors $\mathbf{v}$ and $\mathbf{w}$.

3. $\mathbf{v} = (5,1,-2)$, $\mathbf{w} = (4,-4,3)$

4. $\mathbf{v} = (7,2,-10)$, $\mathbf{w} = (2,6,4)$

5. $\mathbf{v} = (2,1,4)$, $\mathbf{w} = (1,-2,0)$

6. $\mathbf{v} = (4,2,-1)$, $\mathbf{w} = (8,4,-2)$

7. $\mathbf{v} = -\mathbf{i} + 2\mathbf{j} + \mathbf{k}$, $\mathbf{w} = -3\mathbf{i} + 6\mathbf{j} + 3\mathbf{k}$

8. $\mathbf{v} = \mathbf{i}$, $\mathbf{w} = 3\mathbf{i} + 2\mathbf{j} + 4\mathbf{k}$

9. Let $\mathbf{v} = (8,4,3)$ and $\mathbf{w} = (-2,1,4)$. Is $\mathbf{v} \perp \mathbf{w}$? Justify your answer.

10. Let $\mathbf{v} = (6,0,4)$ and $\mathbf{w} = (0,2,-1)$. Is $\mathbf{v} \perp \mathbf{w}$? Justify your answer.

11. For $\mathbf{v}$, $\mathbf{w}$ from Exercise 5, verify the Cauchy-Schwarz Inequality $|\mathbf{v} \cdot \mathbf{w}| \le \|\mathbf{v}\|\,\|\mathbf{w}\|$.

12. For $\mathbf{v}$, $\mathbf{w}$ from Exercise 6, verify the Cauchy-Schwarz Inequality $|\mathbf{v} \cdot \mathbf{w}| \le \|\mathbf{v}\|\,\|\mathbf{w}\|$.

13. For $\mathbf{v}$, $\mathbf{w}$ from Exercise 5, verify the Triangle Inequality $\|\mathbf{v} + \mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\|$.

14. For $\mathbf{v}$, $\mathbf{w}$ from Exercise 6, verify the Triangle Inequality $\|\mathbf{v} + \mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\|$.

B

Note: Consider only vectors in $\mathbb{R}^3$ for Exercises 15-25.

15. Prove Theorem 1.9(a).

16. Prove Theorem 1.9(b).

17. Prove Theorem 1.9(c).

18. Prove Theorem 1.9(d).

19. Prove Theorem 1.9(e).

20. Prove Theorem 1.10(a).

21. Prove or give a counterexample: If $\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot \mathbf{w}$, then $\mathbf{v} = \mathbf{w}$.

C

22. Prove or give a counterexample: If $\mathbf{v} \cdot \mathbf{w} = 0$ for all $\mathbf{v}$, then $\mathbf{w} = \mathbf{0}$.

23. Prove or give a counterexample: If $\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot \mathbf{w}$ for all $\mathbf{u}$, then $\mathbf{v} = \mathbf{w}$.

24. Prove that $\big|\,\|\mathbf{v}\| - \|\mathbf{w}\|\,\big| \le \|\mathbf{v} - \mathbf{w}\|$ for all $\mathbf{v}$, $\mathbf{w}$.

25. For nonzero vectors $\mathbf{v}$ and $\mathbf{w}$, the *projection* of $\mathbf{v}$ onto $\mathbf{w}$ (sometimes written as $proj_{\mathbf{w}}\mathbf{v}$) is the vector $\mathbf{u}$ along the same line L as $\mathbf{w}$ whose terminal point is obtained by dropping a perpendicular line from the terminal point of $\mathbf{v}$ to L (see Figure 1.3.5). Show that

$$\|\mathbf{u}\| = \frac{|\mathbf{v} \cdot \mathbf{w}|}{\|\mathbf{w}\|}.$$

(Hint: Consider the angle between $\mathbf{v}$ *and* $\mathbf{w}$*.)*

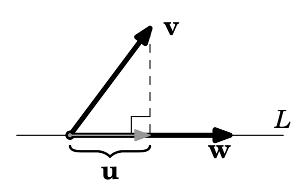


Figure 1.3.5

26. Let α, β, and γ be the angles between a nonzero vector $\mathbf{v}$ in $\mathbb{R}^3$ and the vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$, respectively. Show that $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$.
(Note: α, β, γ are often called the *direction angles* of $\mathbf{v}$, and $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are called the *direction cosines*.)

1.4 Cross Product

In Section 1.3 we defined the dot product, which gave a way of multiplying two vectors. The resulting product, however, was a scalar, not a vector. In this section we will define a product of two vectors that does result in another vector. This product, called the *cross product*, is only defined for vectors in $\mathbb{R}^3$. The definition may appear strange and lacking motivation, but we will see the geometric basis for it shortly.

Definition 1.8. Let $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$ be vectors in $\mathbb{R}^3$. The **cross product** of $\mathbf{v}$ and $\mathbf{w}$, denoted by $\mathbf{v} \times \mathbf{w}$, is the vector in $\mathbb{R}^3$ given by:

$$\mathbf{v} \times \mathbf{w} = (v_2w_3 - v_3w_2, v_3w_1 - v_1w_3, v_1w_2 - v_2w_1) \tag{1.10}$$

Example 1.7. Find $\mathbf{i} \times \mathbf{j}$.

Solution: Since $\mathbf{i} = (1,0,0)$ and $\mathbf{j} = (0,1,0)$, then

$$\begin{aligned}\mathbf{i} \times \mathbf{j} &= ((0)(0) - (0)(1), (0)(0) - (1)(0), (1)(1) - (0)(0))\\ &= (0,0,1)\\ &= \mathbf{k}\end{aligned}$$

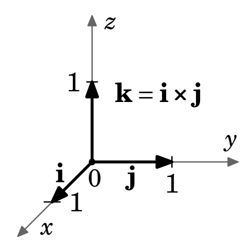


Figure 1.4.1

Similarly it can be shown that $\mathbf{j} \times \mathbf{k} = \mathbf{i}$ and $\mathbf{k} \times \mathbf{i} = \mathbf{j}$.

In the above example, the cross product of the given vectors was perpendicular to both those vectors. It turns out that this will always be the case.

Theorem 1.11. If the cross product $\mathbf{v} \times \mathbf{w}$ of two nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ is also a nonzero vector, then it is perpendicular to both $\mathbf{v}$ and $\mathbf{w}$.

Proof: We will show that $(\mathbf{v} \times \mathbf{w}) \cdot \mathbf{v} = 0$:

$$\begin{aligned}(\mathbf{v} \times \mathbf{w}) \cdot \mathbf{v} &= (v_2w_3 - v_3w_2, v_3w_1 - v_1w_3, v_1w_2 - v_2w_1) \cdot (v_1, v_2, v_3)\\ &= v_2w_3v_1 - v_3w_2v_1 + v_3w_1v_2 - v_1w_3v_2 + v_1w_2v_3 - v_2w_1v_3\\ &= v_1v_2w_3 - v_1v_2w_3 + w_1v_2v_3 - w_1v_2v_3 + v_1w_2v_3 - v_1w_2v_3\\ &= 0 \text{ , after rearranging the terms.}\end{aligned}$$

$\therefore \mathbf{v} \times \mathbf{w} \perp \mathbf{v}$ by Corollary 1.7.
The proof that $\mathbf{v} \times \mathbf{w} \perp \mathbf{w}$ is similar. **QED**

As a consequence of the above theorem and Theorem 1.9, we have the following:

Corollary 1.12. If the cross product $\mathbf{v} \times \mathbf{w}$ of two nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ is also a nonzero vector, then it is perpendicular to the span of $\mathbf{v}$ and $\mathbf{w}$.

The span of any two nonzero, nonparallel vectors $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$ is a plane P, so the above corollary shows that $\mathbf{v}\times\mathbf{w}$ is perpendicular to that plane. As shown in Figure 1.4.2, there are two possible directions for $\mathbf{v}\times\mathbf{w}$, one the opposite of the other. It turns out (see Appendix B) that the direction of $\mathbf{v}\times\mathbf{w}$ is given by the *right-hand rule*, that is, the vectors $\mathbf{v}$, $\mathbf{w}$, $\mathbf{v}\times\mathbf{w}$ form a right-handed system. Recall from Section 1.1 that this means that you can point your thumb upwards in the direction of $\mathbf{v}\times\mathbf{w}$ while rotating $\mathbf{v}$ towards $\mathbf{w}$ with the remaining four fingers.

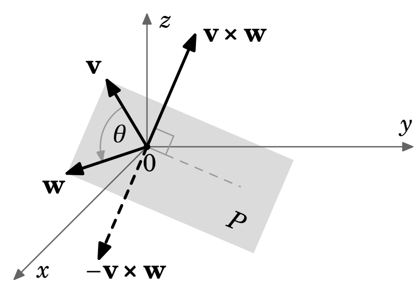


Figure 1.4.2 Direction of $\mathbf{v}\times\mathbf{w}$

We will now derive a formula for the magnitude of $\mathbf{v}\times\mathbf{w}$, for nonzero vectors $\mathbf{v}$, $\mathbf{w}$:

$$\begin{aligned}\|\mathbf{v}\times\mathbf{w}\|^2 &= (v_2w_3-v_3w_2)^2+(v_3w_1-v_1w_3)^2+(v_1w_2-v_2w_1)^2\\ &= v_2^2w_3^2-2v_2w_2v_3w_3+v_3^2w_2^2+v_3^2w_1^2-2v_1w_1v_3w_3+v_1^2w_3^2+v_1^2w_2^2-2v_1w_1v_2w_2+v_2^2w_1^2\\ &= v_1^2(w_2^2+w_3^2)+v_2^2(w_1^2+w_3^2)+v_3^2(w_1^2+w_2^2)-2(v_1w_1v_2w_2+v_1w_1v_3w_3+v_2w_2v_3w_3)\end{aligned}$$

and now adding and subtracting $v_1^2w_1^2$, $v_2^2w_2^2$, and $v_3^2w_3^2$ on the right side gives

$$\begin{aligned}&= v_1^2(w_1^2+w_2^2+w_3^2)+v_2^2(w_1^2+w_2^2+w_3^2)+v_3^2(w_1^2+w_2^2+w_3^2)\\ &\quad-(v_1^2w_1^2+v_2^2w_2^2+v_3^2w_3^2+2(v_1w_1v_2w_2+v_1w_1v_3w_3+v_2w_2v_3w_3))\\ &= (v_1^2+v_2^2+v_3^2)(w_1^2+w_2^2+w_3^2)\\ &\quad-((v_1w_1)^2+(v_2w_2)^2+(v_3w_3)^2+2(v_1w_1)(v_2w_2)+2(v_1w_1)(v_3w_3)+2(v_2w_2)(v_3w_3))\end{aligned}$$

so using $(a+b+c)^2=a^2+b^2+c^2+2ab+2ac+2bc$ for the subtracted term gives

$$\begin{aligned}&= (v_1^2+v_2^2+v_3^2)(w_1^2+w_2^2+w_3^2)-(v_1w_1+v_2w_2+v_3w_3)^2\\ &= \|\mathbf{v}\|^2\,\|\mathbf{w}\|^2-(\mathbf{v}\cdot\mathbf{w})^2\\ &= \|\mathbf{v}\|^2\,\|\mathbf{w}\|^2\left(1-\frac{(\mathbf{v}\cdot\mathbf{w})^2}{\|\mathbf{v}\|^2\,\|\mathbf{w}\|^2}\right)\text{, since }\|\mathbf{v}\|>0\text{ and }\|\mathbf{w}\|>0\text{, so by Theorem 1.6}\\ &= \|\mathbf{v}\|^2\,\|\mathbf{w}\|^2(1-\cos^2\theta)\text{, where }\theta\text{ is the angle between }\mathbf{v}\text{ and }\mathbf{w}\text{, so}\end{aligned}$$

$\|\mathbf{v}\times\mathbf{w}\|^2=\|\mathbf{v}\|^2\,\|\mathbf{w}\|^2\sin^2\theta$, and since $0^\circ\le\theta\le180^\circ$, then $\sin\theta\ge0$, so we have:

If θ is the angle between nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$, then

$$\|\mathbf{v}\times\mathbf{w}\| = \|\mathbf{v}\|\,\|\mathbf{w}\|\,\sin\theta \tag{1.11}$$

It may seem strange to bother with the above formula, when the magnitude of the cross product can be calculated directly, like for any other vector. The formula is more useful for its applications in geometry, as in the following example.

Example 1.8. Let $\triangle PQR$ and $PQRS$ be a triangle and parallelogram, respectively, as shown in Figure 1.4.3.



Figure 1.4.3

Think of the triangle as existing in $\mathbb{R}^3$, and identify the sides QR and QP with vectors $\mathbf{v}$ and $\mathbf{w}$, respectively, in $\mathbb{R}^3$. Let θ be the angle between $\mathbf{v}$ and $\mathbf{w}$. The area A_{PQR} of $\triangle PQR$ is $\frac{1}{2}bh$, where b is the base of the triangle and h is the height. So we see that

$$b = \|\mathbf{v}\| \quad \text{and} \quad h = \|\mathbf{w}\|\sin\theta$$

$$\begin{aligned} A_{PQR} &= \frac{1}{2}\|\mathbf{v}\|\,\|\mathbf{w}\|\sin\theta \\ &= \frac{1}{2}\|\mathbf{v}\times\mathbf{w}\| \end{aligned}$$

So since the area A_{PQRS} of the parallelogram $PQRS$ is twice the area of the triangle $\triangle PQR$, then

$$A_{PQRS} = \|\mathbf{v}\|\,\|\mathbf{w}\|\sin\theta$$

By the discussion in Example 1.8, we have proved the following theorem:

Theorem 1.13. Area of triangles and parallelograms

(a) The area A of a triangle with adjacent sides $\mathbf{v}$, $\mathbf{w}$ (as vectors in $\mathbb{R}^3$) is:

$$A = \frac{1}{2}\|\mathbf{v}\times\mathbf{w}\|$$

(b) The area A of a parallelogram with adjacent sides $\mathbf{v}$, $\mathbf{w}$ (as vectors in $\mathbb{R}^3$) is:

$$A = \|\mathbf{v}\times\mathbf{w}\|$$

It may seem at first glance that since the formulas derived in Example 1.8 were for the adjacent sides QP and QR only, then the more general statements in Theorem 1.13 that the formulas hold for *any* adjacent sides are not justified. We would get a different *formula* for the area if we had picked PQ and PR as the adjacent sides, but it can be shown (see Exercise 26) that the different formulas would yield the same value, so the choice of adjacent sides indeed does not matter, and Theorem 1.13 is valid.

Theorem 1.13 makes it simpler to calculate the area of a triangle in 3-dimensional space than by using traditional geometric methods.

Example 1.9. Calculate the area of the triangle $\triangle PQR$, where $P = (2,4,-7)$, $Q = (3,7,18)$, and $R = (-5,12,8)$.

Solution: Let $\mathbf{v} = \overrightarrow{PQ}$ and $\mathbf{w} = \overrightarrow{PR}$, as in Figure 1.4.4. Then $\mathbf{v} = (3,7,18)-(2,4,-7) = (1,3,25)$ and $\mathbf{w} = (-5,12,8)-(2,4,-7) = (-7,8,15)$, so the area A of the triangle $\triangle PQR$ is

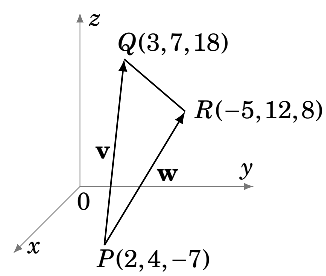


Figure 1.4.4

$$
\begin{aligned}
A &= \frac{1}{2}\,\|\mathbf{v}\times\mathbf{w}\| = \frac{1}{2}\,\|(1,3,25)\times(-7,8,15)\|\\
&= \frac{1}{2}\,\big\|((3)(15)-(25)(8),(25)(-7)-(1)(15),(1)(8)-(3)(-7))\big\|\\
&= \frac{1}{2}\,\big\|(-155,-190,29)\big\|\\
&= \frac{1}{2}\,\sqrt{(-155)^2+(-190)^2+29^2} = \frac{1}{2}\sqrt{60966}\\
A &\approx 123.46
\end{aligned}
$$

Example 1.10. Calculate the area of the parallelogram $PQRS$, where $P = (1,1)$, $Q = (2,3)$, $R = (5,4)$, and $S = (4,2)$.

Solution: Let $\mathbf{v} = \overrightarrow{SP}$ and $\mathbf{w} = \overrightarrow{SR}$, as in Figure 1.4.5. Then $\mathbf{v} = (1,1)-(4,2) = (-3,-1)$ and $\mathbf{w} = (5,4)-(4,2) = (1,2)$. But these are vectors in $\mathbb{R}^2$, and the cross product is only defined for vectors in $\mathbb{R}^3$. However, $\mathbb{R}^2$ can be thought of as the subset of $\mathbb{R}^3$ such that the z-coordinate is always 0. So we can write $\mathbf{v} = (-3,-1,0)$ and $\mathbf{w} = (1,2,0)$. Then the area A of $PQRS$ is

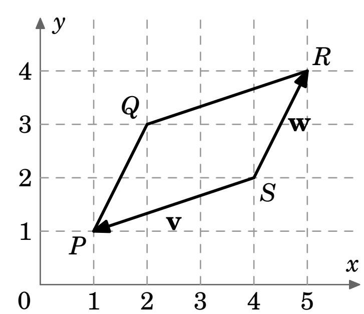


Figure 1.4.5

$$
\begin{aligned}
A &= \|\mathbf{v}\times\mathbf{w}\| = \big\|(-3,-1,0)\times(1,2,0)\big\|\\
&= \big\|((-1)(0)-(0)(2),(0)(1)-(-3)(0),(-3)(2)-(-1)(1))\big\|\\
&= \big\|(0,0,-5)\big\|\\
A &= 5
\end{aligned}
$$

The following theorem summarizes the basic properties of the cross product.

Theorem 1.14. For any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$, and scalar k, we have

(a) $\mathbf{v} \times \mathbf{w} = -\mathbf{w} \times \mathbf{v}$ — Anticommutative Law

(b) $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w}$ — Distributive Law

(c) $(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = \mathbf{u} \times \mathbf{w} + \mathbf{v} \times \mathbf{w}$ — Distributive Law

(d) $(k\mathbf{v}) \times \mathbf{w} = \mathbf{v} \times (k\mathbf{w}) = k(\mathbf{v} \times \mathbf{w})$ — Associative Law

(e) $\mathbf{v} \times \mathbf{0} = \mathbf{0} = \mathbf{0} \times \mathbf{v}$

(f) $\mathbf{v} \times \mathbf{v} = \mathbf{0}$

(g) $\mathbf{v} \times \mathbf{w} = \mathbf{0}$ if and only if $\mathbf{v} \parallel \mathbf{w}$

Proof: The proofs of properties (b)-(f) are straightforward. We will prove parts (a) and (g) and leave the rest to the reader as exercises.

(a) By the definition of the cross product and scalar multiplication, we have:

$$\begin{aligned}
\mathbf{v} \times \mathbf{w} &= (v_2w_3 - v_3w_2, v_3w_1 - v_1w_3, v_1w_2 - v_2w_1) \\
&= -(v_3w_2 - v_2w_3, v_1w_3 - v_3w_1, v_2w_1 - v_1w_2) \\
&= -(w_2v_3 - w_3v_2, w_3v_1 - w_1v_3, w_1v_2 - w_2v_1) \\
&= -\mathbf{w} \times \mathbf{v}
\end{aligned}$$

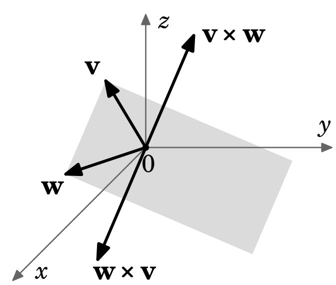


Figure 1.4.6

Note that this says that $\mathbf{v} \times \mathbf{w}$ and $\mathbf{w} \times \mathbf{v}$ have the same magnitude but opposite direction (see Figure 1.4.6).

(g) If either $\mathbf{v}$ or $\mathbf{w}$ is $\mathbf{0}$ then $\mathbf{v} \times \mathbf{w} = \mathbf{0}$ by part (e), and either $\mathbf{v} = \mathbf{0} = 0\mathbf{w}$ or $\mathbf{w} = \mathbf{0} = 0\mathbf{v}$, so $\mathbf{v}$ and $\mathbf{w}$ are scalar multiples, i.e. they are parallel.

If both $\mathbf{v}$ and $\mathbf{w}$ are nonzero, and θ is the angle between them, then by formula (1.11), $\mathbf{v} \times \mathbf{w} = \mathbf{0}$ if and only if $\|\mathbf{v}\|\,\|\mathbf{w}\| \sin\theta = 0$, which is true if and only if $\sin\theta = 0$ (since $\|\mathbf{v}\| > 0$ and $\|\mathbf{w}\| > 0$). So since $0° \le \theta \le 180°$, then $\sin\theta = 0$ if and only if $\theta = 0°$ or $180°$. But the angle between $\mathbf{v}$ and $\mathbf{w}$ is $0°$ or $180°$ if and only if $\mathbf{v} \parallel \mathbf{w}$. **QED**

Example 1.11. Adding to Example 1.7, we have

$$\mathbf{i} \times \mathbf{j} = \mathbf{k} \qquad \mathbf{j} \times \mathbf{k} = \mathbf{i} \qquad \mathbf{k} \times \mathbf{i} = \mathbf{j}$$
$$\mathbf{j} \times \mathbf{i} = -\mathbf{k} \qquad \mathbf{k} \times \mathbf{j} = -\mathbf{i} \qquad \mathbf{i} \times \mathbf{k} = -\mathbf{j}$$
$$\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = \mathbf{0}$$

Recall from geometry that a *parallelepiped* is a 3-dimensional solid with 6 faces, all of which are parallelograms.[6]

[6] An equivalent definition of a parallelepiped is: the collection of all scalar combinations $k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + k_3\mathbf{v}_3$ of some vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ in $\mathbb{R}^3$, where $0 \le k_1, k_2, k_3 \le 1$.

Example 1.12. *Volume of a parallelepiped:* Let the vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$ represent adjacent sides of a parallelepiped P, with $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ forming a right-handed system, as in Figure 1.4.7. Show that the volume of P is the *scalar triple product* $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$.

Solution: Recall that the volume vol(P) of a parallelepiped P is the area A of the base parallelogram times the height h. By Theorem 1.13(b), the area A of the base parallelogram is $\|\mathbf{v} \times \mathbf{w}\|$. And we can see that since $\mathbf{v} \times \mathbf{w}$ is perpendicular to the base parallelogram determined by $\mathbf{v}$ and $\mathbf{w}$, then the height h is $\|\mathbf{u}\| \cos\theta$, where θ is the angle between $\mathbf{u}$ and $\mathbf{v} \times \mathbf{w}$. By Theorem 1.6 we know that

$$\cos\theta = \frac{\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})}{\|\mathbf{u}\| \, \|\mathbf{v} \times \mathbf{w}\|} . \text{ Hence,}$$

$$\begin{aligned} \text{vol}(P) &= A\,h \\ &= \|\mathbf{v} \times \mathbf{w}\| \, \frac{\|\mathbf{u}\| \, \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})}{\|\mathbf{u}\| \, \|\mathbf{v} \times \mathbf{w}\|} \\ &= \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) \end{aligned}$$

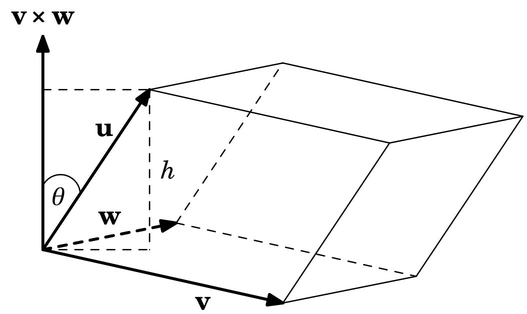


Figure 1.4.7 Parallelepiped P

In Example 1.12 the height h of the parallelepiped is $\|\mathbf{u}\| \cos\theta$, and not $-\|\mathbf{u}\| \cos\theta$, because the vector $\mathbf{u}$ is on the same side of the base parallelogram's plane as the vector $\mathbf{v} \times \mathbf{w}$ (so that $\cos\theta > 0$). Since the volume is the same no matter which base and height we use, then repeating the same steps using the base determined by $\mathbf{u}$ and $\mathbf{v}$ (since $\mathbf{w}$ is on the same side of that base's plane as $\mathbf{u} \times \mathbf{v}$), the volume is $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$. Repeating this with the base determined by $\mathbf{w}$ and $\mathbf{u}$, we have the following result:

For any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$,

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot (\mathbf{w} \times \mathbf{u}) \tag{1.12}$$

(Note that the equalities hold trivially if any of the vectors are $\mathbf{0}$.)

Since $\mathbf{v} \times \mathbf{w} = -\mathbf{w} \times \mathbf{v}$ for any vectors $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$, then picking the wrong order for the three adjacent sides in the scalar triple product in formula (1.12) will give you the negative of the volume of the parallelepiped. So taking the absolute value of the scalar triple product for *any* order of the three adjacent sides will *always* give the volume:

Theorem 1.15. If vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$ represent any three adjacent sides of a parallelepiped, then the volume of the parallelepiped is $|\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})|$.

Another type of triple product is the *vector triple product* $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$. The proof of the following theorem is left as an exercise for the reader:

Theorem 1.16. For any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$,

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w} \tag{1.13}$$

An examination of the formula in Theorem 1.16 gives some idea of the geometry of the vector triple product. By the right side of formula (1.13), we see that $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ is a scalar combination of $\mathbf{v}$ and $\mathbf{w}$, and hence lies in the plane containing $\mathbf{v}$ and $\mathbf{w}$ (i.e. $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$, $\mathbf{v}$ and $\mathbf{w}$ are **coplanar**). This makes sense since, by Theorem 1.11, $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ is perpendicular to both $\mathbf{u}$ and $\mathbf{v} \times \mathbf{w}$. In particular, being perpendicular to $\mathbf{v} \times \mathbf{w}$ means that $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ lies in the plane containing $\mathbf{v}$ and $\mathbf{w}$, since that plane is itself perpendicular to $\mathbf{v} \times \mathbf{w}$. But then how is $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ also perpendicular to $\mathbf{u}$, which could be any vector? The following example may help to see how this works.

Example 1.13. Find $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ for $\mathbf{u} = (1,2,4)$, $\mathbf{v} = (2,2,0)$, $\mathbf{w} = (1,3,0)$.

Solution: Since $\mathbf{u} \cdot \mathbf{v} = 6$ and $\mathbf{u} \cdot \mathbf{w} = 7$, then

$$\begin{aligned}\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) &= (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}\\ &= 7(2,2,0) - 6(1,3,0) = (14,14,0) - (6,18,0)\\ &= (8,-4,0)\end{aligned}$$

Note that $\mathbf{v}$ and $\mathbf{w}$ lie in the xy-plane, and that $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ also lies in that plane. Also, $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ is perpendicular to both $\mathbf{u}$ and $\mathbf{v} \times \mathbf{w} = (0,0,4)$ (see Figure 1.4.8).

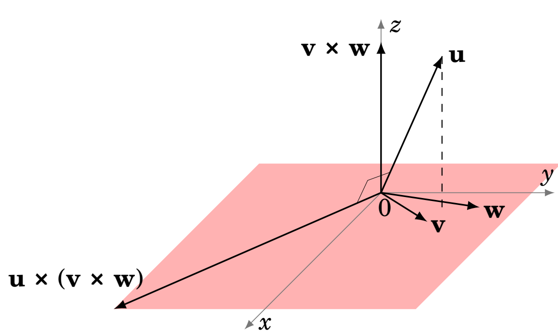


Figure 1.4.8

For vectors $\mathbf{v} = v_1\,\mathbf{i} + v_2\,\mathbf{j} + v_3\,\mathbf{k}$ and $\mathbf{w} = w_1\,\mathbf{i} + w_2\,\mathbf{j} + w_3\,\mathbf{k}$ in component form, the cross product is written as: $\mathbf{v} \times \mathbf{w} = (v_2 w_3 - v_3 w_2)\mathbf{i} + (v_3 w_1 - v_1 w_3)\mathbf{j} + (v_1 w_2 - v_2 w_1)\mathbf{k}$. It is often easier to use the component form for the cross product, because it can be represented as a *determinant*. We will not go too deeply into the theory of determinants[7]; we will just cover what is essential for our purposes.

[7] See ANTON and RORRES for a fuller development.

A **2 × 2 matrix** is an array of two rows and two columns of scalars, written as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{or} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where a, b, c, d are scalars. The **determinant** of such a matrix, written as

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} \quad \text{or} \quad \det \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

is the scalar defined by the following formula:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

It may help to remember this formula as being the product of the scalars on the downward diagonal minus the product of the scalars on the upward diagonal.

Example 1.14.

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = (1)(4) - (2)(3) = 4 - 6 = -2$$

A **3 × 3 matrix** is an array of three rows and three columns of scalars, written as

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} \quad \text{or} \quad \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix},$$

and its determinant is given by the formula:

$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix} \tag{1.14}$$

One way to remember the above formula is the following: multiply each scalar in the first row by the determinant of the 2×2 matrix that remains after removing the row and column that contain that scalar, then sum those products up, putting alternating plus and minus signs in front of each (starting with a plus).

Example 1.15.

$$\begin{vmatrix} 1 & 0 & 2 \\ 4 & -1 & 3 \\ 1 & 0 & 2 \end{vmatrix} = 1 \begin{vmatrix} -1 & 3 \\ 0 & 2 \end{vmatrix} - 0 \begin{vmatrix} 4 & 3 \\ 1 & 2 \end{vmatrix} + 2 \begin{vmatrix} 4 & -1 \\ 1 & 0 \end{vmatrix} = 1(-2-0) - 0(8-3) + 2(0+1) = 0$$

We defined the determinant as a scalar, derived from algebraic operations on scalar entries in a matrix. However, if we put three *vectors* in the first row of a 3×3 matrix, then the definition still makes sense, since we would be performing scalar multiplication on those three vectors (they would be multiplied by the 2×2 scalar determinants as before). This gives us a determinant that is now a vector, and lets us write the cross product of $\mathbf{v} = v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}$ and $\mathbf{w} = w_1\,\mathbf{i}+w_2\,\mathbf{j}+w_3\,\mathbf{k}$ as a determinant:

$$\mathbf{v}\times\mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = \begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix}\mathbf{i} - \begin{vmatrix} v_1 & v_3 \\ w_1 & w_3 \end{vmatrix}\mathbf{j} + \begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix}\mathbf{k}$$
$$= (v_2w_3 - v_3w_2)\mathbf{i} + (v_3w_1 - v_1w_3)\mathbf{j} + (v_1w_2 - v_2w_1)\mathbf{k}$$

Example 1.16. Let $\mathbf{v} = 4\,\mathbf{i}-\mathbf{j}+3\,\mathbf{k}$ and $\mathbf{w} = \mathbf{i}+2\,\mathbf{k}$. Then

$$\mathbf{v}\times\mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 4 & -1 & 3 \\ 1 & 0 & 2 \end{vmatrix} = \begin{vmatrix} -1 & 3 \\ 0 & 2 \end{vmatrix}\mathbf{i} - \begin{vmatrix} 4 & 3 \\ 1 & 2 \end{vmatrix}\mathbf{j} + \begin{vmatrix} 4 & -1 \\ 1 & 0 \end{vmatrix}\mathbf{k} = -2\,\mathbf{i} - 5\,\mathbf{j} + \mathbf{k}$$

The scalar triple product can also be written as a determinant. In fact, by Example 1.12, the following theorem provides an alternate definition of the determinant of a 3×3 matrix as the volume of a parallelepiped whose adjacent sides are the rows of the matrix and form a right-handed system (a left-handed system would give the negative volume).

Theorem 1.17. For any vectors $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$, $\mathbf{w} = (w_1, w_2, w_3)$ in $\mathbb{R}^3$:

$$\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \tag{1.15}$$

Example 1.17. Find the volume of the parallelepiped with adjacent sides $\mathbf{u} = (2,1,3)$, $\mathbf{v} = (-1,3,2)$, $\mathbf{w} = (1,1,-2)$ (see Figure 1.4.9).

Solution: By Theorem 1.15, the volume vol(P) of the parallelepiped P is the absolute value of the scalar triple product of the three adjacent sides (in any order). By Theorem 1.17,

$$\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w}) = \begin{vmatrix} 2 & 1 & 3 \\ -1 & 3 & 2 \\ 1 & 1 & -2 \end{vmatrix}$$
$$= 2\begin{vmatrix} 3 & 2 \\ 1 & -2 \end{vmatrix} - 1\begin{vmatrix} -1 & 2 \\ 1 & -2 \end{vmatrix} + 3\begin{vmatrix} -1 & 3 \\ 1 & 1 \end{vmatrix}$$
$$= 2(-8) - 1(0) + 3(-4) = -28, \text{ so}$$
$$\text{vol}(P) = |-28| = 28.$$

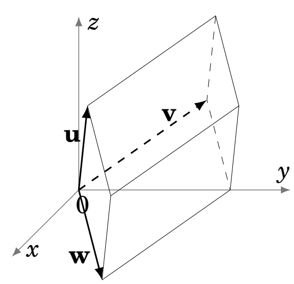


Figure 1.4.9 P

Interchanging the dot and cross products can be useful in proving vector identities:

Example 1.18. Prove: $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{w} \times \mathbf{z}) = \begin{vmatrix} \mathbf{u}\cdot\mathbf{w} & \mathbf{u}\cdot\mathbf{z} \\ \mathbf{v}\cdot\mathbf{w} & \mathbf{v}\cdot\mathbf{z} \end{vmatrix}$ for all vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, $\mathbf{z}$ in $\mathbb{R}^3$.

Solution: Let $\mathbf{x} = \mathbf{u} \times \mathbf{v}$. Then

$$\begin{aligned}
(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{w} \times \mathbf{z}) &= \mathbf{x} \cdot (\mathbf{w} \times \mathbf{z}) \\
&= \mathbf{w} \cdot (\mathbf{z} \times \mathbf{x}) \quad \text{(by formula (1.12))} \\
&= \mathbf{w} \cdot (\mathbf{z} \times (\mathbf{u} \times \mathbf{v})) \\
&= \mathbf{w} \cdot ((\mathbf{z}\cdot\mathbf{v})\mathbf{u} - (\mathbf{z}\cdot\mathbf{u})\mathbf{v}) \quad \text{(by Theorem 1.16)} \\
&= (\mathbf{z}\cdot\mathbf{v})(\mathbf{w}\cdot\mathbf{u}) - (\mathbf{z}\cdot\mathbf{u})(\mathbf{w}\cdot\mathbf{v}) \\
&= (\mathbf{u}\cdot\mathbf{w})(\mathbf{v}\cdot\mathbf{z}) - (\mathbf{u}\cdot\mathbf{z})(\mathbf{v}\cdot\mathbf{w}) \quad \text{(by commutativity of the dot product).} \\
&= \begin{vmatrix} \mathbf{u}\cdot\mathbf{w} & \mathbf{u}\cdot\mathbf{z} \\ \mathbf{v}\cdot\mathbf{w} & \mathbf{v}\cdot\mathbf{z} \end{vmatrix}
\end{aligned}$$

Exercises

A

For Exercises 1-6, calculate $\mathbf{v} \times \mathbf{w}$.

1. $\mathbf{v} = (5,1,-2)$, $\mathbf{w} = (4,-4,3)$

2. $\mathbf{v} = (7,2,-10)$, $\mathbf{w} = (2,6,4)$

3. $\mathbf{v} = (2,1,4)$, $\mathbf{w} = (1,-2,0)$

4. $\mathbf{v} = (1,3,2)$, $\mathbf{w} = (7,2,-10)$

5. $\mathbf{v} = -\mathbf{i} + 2\mathbf{j} + \mathbf{k}$, $\mathbf{w} = -3\mathbf{i} + 6\mathbf{j} + 3\mathbf{k}$

6. $\mathbf{v} = \mathbf{i}$, $\mathbf{w} = 3\mathbf{i} + 2\mathbf{j} + 4\mathbf{k}$

For Exercises 7-8, calculate the area of the triangle $\triangle PQR$.

7. $P = (5,1,-2)$, $Q = (4,-4,3)$, $R = (2,4,0)$

8. $P = (4,0,2)$, $Q = (2,1,5)$, $R = (-1,0,-1)$

For Exercises 9-10, calculate the area of the parallelogram $PQRS$.

9. $P = (2,1,3)$, $Q = (1,4,5)$, $R = (2,5,3)$, $S = (3,2,1)$

10. $P = (-2,-2)$, $Q = (1,4)$, $R = (6,6)$, $S = (3,0)$

For Exercises 11-12, find the volume of the parallelepiped with adjacent sides $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$.

11. $\mathbf{u} = (1,1,3)$, $\mathbf{v} = (2,1,4)$, $\mathbf{w} = (5,1,-2)$

12. $\mathbf{u} = (1,3,2)$, $\mathbf{v} = (7,2,-10)$, $\mathbf{w} = (1,0,1)$

For Exercises 13-14, calculate $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$ and $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$.

13. $\mathbf{u} = (1,1,1)$, $\mathbf{v} = (3,0,2)$, $\mathbf{w} = (2,2,2)$

14. $\mathbf{u} = (1,0,2)$, $\mathbf{v} = (-1,0,3)$, $\mathbf{w} = (2,0,-2)$

15. Calculate $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{w} \times \mathbf{z})$ for $\mathbf{u} = (1,1,1)$, $\mathbf{v} = (3,0,2)$, $\mathbf{w} = (2,2,2)$, $\mathbf{z} = (2,1,4)$.

B

16. If $\mathbf{v}$ and $\mathbf{w}$ are unit vectors in $\mathbb{R}^3$, under what condition(s) would $\mathbf{v} \times \mathbf{w}$ also be a unit vector in $\mathbb{R}^3$? Justify your answer.

17. Show that if $\mathbf{v} \times \mathbf{w} = \mathbf{0}$ for all $\mathbf{w}$ in $\mathbb{R}^3$, then $\mathbf{v} = \mathbf{0}$.

18. Prove Theorem 1.14(b).

19. Prove Theorem 1.14(c).

20. Prove Theorem 1.14(d).

21. Prove Theorem 1.14(e).

22. Prove Theorem 1.14(f).

23. Prove Theorem 1.16.

24. Prove Theorem 1.17. (*Hint: Expand both sides of the equation.*)

25. Prove the following for all vectors $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$:

(a) $\|\mathbf{v} \times \mathbf{w}\|^2 + |\mathbf{v} \cdot \mathbf{w}|^2 = \|\mathbf{v}\|^2 \, \|\mathbf{w}\|^2$

(b) If $\mathbf{v} \cdot \mathbf{w} = 0$ and $\mathbf{v} \times \mathbf{w} = \mathbf{0}$, then $\mathbf{v} = \mathbf{0}$ or $\mathbf{w} = \mathbf{0}$.

C

26. Prove that in Example 1.8 the formula for the area of the triangle $\triangle PQR$ yields the same value no matter which two adjacent sides are chosen. To do this, show that $\frac{1}{2}\,\|\mathbf{u} \times (-\mathbf{w})\| = \frac{1}{2}\,\|\mathbf{v} \times \mathbf{w}\|$, where $\mathbf{u} = PR$, $-\mathbf{w} = PQ$, and $\mathbf{v} = QR$, $\mathbf{w} = QP$ as before. Similarly, show that $\frac{1}{2}\,\|(-\mathbf{u}) \times (-\mathbf{v})\| = \frac{1}{2}\,\|\mathbf{v} \times \mathbf{w}\|$, where $-\mathbf{u} = RP$ and $-\mathbf{v} = RQ$.

27. Consider the vector equation $\mathbf{a} \times \mathbf{x} = \mathbf{b}$ in $\mathbb{R}^3$, where $\mathbf{a} \neq \mathbf{0}$. Show that:

(a) $\mathbf{a} \cdot \mathbf{b} = 0$

(b) $\mathbf{x} = \dfrac{\mathbf{b} \times \mathbf{a}}{\|\mathbf{a}\|^2} + k\mathbf{a}$ is a solution to the equation, for any scalar k

28. Prove the *Jacobi identity*: $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) + \mathbf{v} \times (\mathbf{w} \times \mathbf{u}) + \mathbf{w} \times (\mathbf{u} \times \mathbf{v}) = \mathbf{0}$

29. Show that $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ lie in the same plane in $\mathbb{R}^3$ if and only if $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = 0$.

30. For all vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, $\mathbf{z}$ in $\mathbb{R}^3$, show that

$$(\mathbf{u} \times \mathbf{v}) \times (\mathbf{w} \times \mathbf{z}) = (\mathbf{z} \cdot (\mathbf{u} \times \mathbf{v}))\mathbf{w} - (\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}))\mathbf{z}$$

and that

$$(\mathbf{u} \times \mathbf{v}) \times (\mathbf{w} \times \mathbf{z}) = (\mathbf{u} \cdot (\mathbf{w} \times \mathbf{z}))\mathbf{v} - (\mathbf{v} \cdot (\mathbf{w} \times \mathbf{z}))\mathbf{u}$$

Why do both equations make sense geometrically?

1.5 Lines and Planes

Now that we know how to perform some operations on vectors, we can start to deal with some familiar geometric objects, like lines and planes, in the language of vectors. The reason for doing this is simple: using vectors makes it easier to study objects in 3-dimensional Euclidean space. We will first consider lines.

Line through a point, parallel to a vector

Let $P = (x_0, y_0, z_0)$ be a point in $\mathbb{R}^3$, let $\mathbf{v} = (a, b, c)$ be a nonzero vector, and let L be the line through P which is parallel to $\mathbf{v}$ (see Figure 1.5.1).

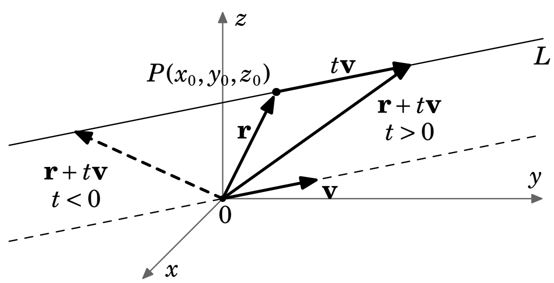


Figure 1.5.1

Let $\mathbf{r} = (x_0, y_0, z_0)$ be the *vector* pointing from the origin to P. Since multiplying the vector $\mathbf{v}$ by a scalar t lengthens or shrinks $\mathbf{v}$ while preserving its direction if $t > 0$, and reversing its direction if $t < 0$, then we see from Figure 1.5.1 that every point on the line L can be obtained by adding the vector $t\mathbf{v}$ to the vector $\mathbf{r}$ for some scalar t. That is, as t varies over all real numbers, the vector $\mathbf{r} + t\mathbf{v}$ will point to every point on L. We can summarize the *vector representation of* L as follows:

For a point $P = (x_0, y_0, z_0)$ and nonzero vector $\mathbf{v}$ in $\mathbb{R}^3$, the line L through P parallel to $\mathbf{v}$ is given by

$$\mathbf{r} + t\mathbf{v}, \quad \text{for } -\infty < t < \infty \tag{1.16}$$

where $\mathbf{r} = (x_0, y_0, z_0)$ is the vector pointing to P.

Note that we used the correspondence between a vector and its terminal point. Since $\mathbf{v} = (a, b, c)$, then the terminal point of the vector $\mathbf{r} + t\mathbf{v}$ is $(x_0 + at, y_0 + bt, z_0 + ct)$. We then get the *parametric representation of* L with the *parameter* t:

For a point $P = (x_0, y_0, z_0)$ and nonzero vector $\mathbf{v} = (a, b, c)$ in $\mathbb{R}^3$, the line L through P parallel to $\mathbf{v}$ consists of all points (x, y, z) given by

$$x = x_0 + at, \quad y = y_0 + bt, \quad z = z_0 + ct, \quad \text{for } -\infty < t < \infty \tag{1.17}$$

Note that in both representations we get the point P on L by letting $t = 0$.

In formula (1.17), if $a \neq 0$, then we can solve for the parameter t: $t = (x - x_0)/a$. We can also solve for t in terms of y and in terms of z if neither b nor c, respectively, is zero: $t = (y - y_0)/b$ and $t = (z - z_0)/c$. These three values all equal the same value t, so we can write the following system of equalities, called the *symmetric representation of L*:

For a point $P = (x_0, y_0, z_0)$ and vector $\mathbf{v} = (a, b, c)$ in $\mathbb{R}^3$ with a, b and c all nonzero, the line L through P parallel to $\mathbf{v}$ consists of all points (x, y, z) given by the equations

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c} \tag{1.18}$$

What if, say, $a = 0$ in the above scenario? We can not divide by zero, but we do know that $x = x_0 + at$, and so $x = x_0 + 0t = x_0$. Then the symmetric representation of L would be:

$$x = x_0, \quad \frac{y - y_0}{b} = \frac{z - z_0}{c} \tag{1.19}$$

Note that this says that the line L lies in the *plane* $x = x_0$, which is parallel to the yz-plane (see Figure 1.5.2). Similar equations can be derived for the cases when $b = 0$ or $c = 0$.

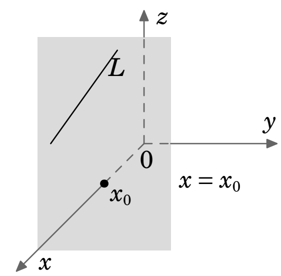


Figure 1.5.2

You may have noticed that the vector representation of L in formula (1.16) is more compact than the parametric and symmetric formulas. That is an advantage of using vector notation. Technically, though, the vector representation gives us the *vectors* whose terminal points make up the line L, not just L itself. So you have to remember to identify the vectors $\mathbf{r} + t\mathbf{v}$ with their terminal points. On the other hand, the parametric representation *always* gives just the points on L and nothing else.

Example 1.19. Write the line L through the point $P = (2,3,5)$ and parallel to the vector $\mathbf{v} = (4,-1,6)$, in the following forms: (a) vector, (b) parametric, (c) symmetric. Lastly: (d) find two points on L distinct from P.

Solution: (a) Let $\mathbf{r} = (2,3,5)$. Then by formula (1.16), L is given by:

$$\mathbf{r} + t\mathbf{v} = (2,3,5) + t(4,-1,6), \quad \text{for } -\infty < t < \infty$$

(b) L consists of the points (x, y, z) such that

$$x = 2 + 4t, \quad y = 3 - t, \quad z = 5 + 6t, \quad \text{for } -\infty < t < \infty$$

(c) L consists of the points (x, y, z) such that

$$\frac{x - 2}{4} = \frac{y - 3}{-1} = \frac{z - 5}{6}$$

(d) Letting $t = 1$ and $t = 2$ in part(b) yields the points $(6,2,11)$ and $(10,1,17)$ on L.

Line through two points

Let $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ be distinct points in $\mathbb{R}^3$, and let L be the line through P_1 and P_2. Let $\mathbf{r}_1 = (x_1, y_1, z_1)$ and $\mathbf{r}_2 = (x_2, y_2, z_2)$ be the vectors pointing to P_1 and P_2, respectively. Then as we can see from Figure 1.5.3, $\mathbf{r}_2 - \mathbf{r}_1$ is the vector from P_1 to P_2. So if we multiply the vector $\mathbf{r}_2 - \mathbf{r}_1$ by a scalar t and add it to the vector $\mathbf{r}_1$, we will get the entire line L as t varies over all real numbers. The following is a summary of the vector, parametric, and symmetric forms for the line L:

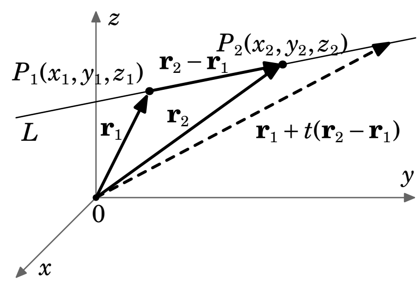


Figure 1.5.3

Let $P_1 = (x_1, y_1, z_1)$, $P_2 = (x_2, y_2, z_2)$ be distinct points in $\mathbb{R}^3$, and let $\mathbf{r}_1 = (x_1, y_1, z_1)$, $\mathbf{r}_2 = (x_2, y_2, z_2)$. Then the line L through P_1 and P_2 has the following representations:

Vector:

$$\mathbf{r}_1 + t(\mathbf{r}_2 - \mathbf{r}_1)\ , \ \text{for } -\infty < t < \infty \tag{1.20}$$

Parametric:

$$x = x_1 + (x_2 - x_1)t, \quad y = y_1 + (y_2 - y_1)t, \quad z = z_1 + (z_2 - z_1)t, \ \text{for } -\infty < t < \infty \tag{1.21}$$

Symmetric:

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1} \quad (\text{if } x_1 \neq x_2,\ y_1 \neq y_2, \text{ and } z_1 \neq z_2) \tag{1.22}$$

Example 1.20. Write the line L through the points $P_1 = (-3, 1, -4)$ and $P_2 = (4, 4, -6)$ in parametric form.

Solution: By formula (1.21), L consists of the points (x, y, z) such that

$$x = -3 + 7t, \quad y = 1 + 3t, \quad z = -4 - 2t, \ \text{for } -\infty < t < \infty$$

Distance between a point and a line

Let L be a line in $\mathbb{R}^3$ in vector form as $\mathbf{r} + t\mathbf{v}$ (for $-\infty < t < \infty$), and let P be a point not on L. The distance d from P to L is the length of the line segment from P to L which is perpendicular to L (see Figure 1.5.4). Pick a point Q on L, and let $\mathbf{w}$ be the vector from Q to P. If θ is the angle between $\mathbf{w}$ and $\mathbf{v}$, then $d = \|\mathbf{w}\| \sin\theta$. So since $\|\mathbf{v} \times \mathbf{w}\| = \|\mathbf{v}\|\,\|\mathbf{w}\| \sin\theta$ and $\mathbf{v} \neq \mathbf{0}$, then:

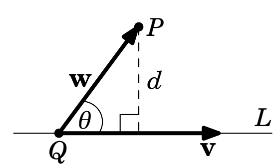


Figure 1.5.4

$$d = \frac{\|\mathbf{v} \times \mathbf{w}\|}{\|\mathbf{v}\|} \tag{1.23}$$

Example 1.21. Find the distance d from the point $P = (1,1,1)$ to the line L in Example 1.20.

Solution: From Example 1.20, we see that we can represent L in vector form as: $\mathbf{r} + t\mathbf{v}$, for $\mathbf{r} = (-3,1,-4)$ and $\mathbf{v} = (7,3,-2)$. Since the point $Q = (-3,1,-4)$ is on L, then for $\mathbf{w} = \overrightarrow{QP} = (1,1,1) - (-3,1,-4) = (4,0,5)$, we have:

$$\mathbf{v} \times \mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 7 & 3 & -2 \\ 4 & 0 & 5 \end{vmatrix} = \begin{vmatrix} 3 & -2 \\ 0 & 5 \end{vmatrix} \mathbf{i} - \begin{vmatrix} 7 & -2 \\ 4 & 5 \end{vmatrix} \mathbf{j} + \begin{vmatrix} 7 & 3 \\ 4 & 0 \end{vmatrix} \mathbf{k} = 15\,\mathbf{i} - 43\,\mathbf{j} - 12\,\mathbf{k} \text{ , so}$$

$$d = \frac{\|\mathbf{v} \times \mathbf{w}\|}{\|\mathbf{v}\|} = \frac{\|15\,\mathbf{i} - 43\,\mathbf{j} - 12\,\mathbf{k}\|}{\|(7,3,-2)\|} = \frac{\sqrt{15^2 + (-43)^2 + (-12)^2}}{\sqrt{7^2 + 3^2 + (-2)^2}} = \frac{\sqrt{2218}}{\sqrt{62}} = 5.98$$

It is clear that two lines L_1 and L_2, represented in vector form as $\mathbf{r}_1 + s\mathbf{v}_1$ and $\mathbf{r}_2 + t\mathbf{v}_2$, respectively, are parallel (denoted as $L_1 \parallel L_2$) if $\mathbf{v}_1$ and $\mathbf{v}_2$ are parallel. Also, L_1 and L_2 are perpendicular (denoted as $L_1 \perp L_2$) if $\mathbf{v}_1$ and $\mathbf{v}_2$ are perpendicular.

In 2-dimensional space, two lines are either identical, parallel, or they intersect. In 3-dimensional space, there is an additional possibility: two lines can be **skew**, that is, they do not intersect but they are not parallel. However, even though they are not parallel, skew lines are on parallel planes (see Figure 1.5.5).

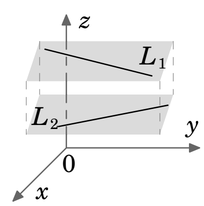


Figure 1.5.5

To determine whether two lines in $\mathbb{R}^3$ intersect, it is often easier to use the parametric representation of the lines. In this case, you should use different parameter variables (usually s and t) for the lines, since the values of the parameters may not be the same at the point of intersection. Setting the two (x,y,z) triples equal will result in a system of 3 equations in 2 unknowns (s and t).

Example 1.22. Find the point of intersection (if any) of the following lines:

$$\frac{x+1}{3} = \frac{y-2}{2} = \frac{z-1}{-1} \quad \text{and} \quad x+3 = \frac{y-8}{-3} = \frac{z+3}{2}$$

Solution: First we write the lines in parametric form, with parameters s and t:

$$x = -1+3s, \quad y = 2+2s, \quad z = 1-s \quad \text{and} \quad x = -3+t, \quad y = 8-3t, \quad z = -3+2t$$

The lines intersect when $(-1+3s, 2+2s, 1-s) = (-3+t, 8-3t, -3+2t)$ for some s, t:

$$-1+3s = -3+t: \Rightarrow t = 2+3s$$
$$2+2s = 8-3t: \Rightarrow 2+2s = 8-3(2+3s) = 2-9s \Rightarrow 2s = -9s \Rightarrow s = 0 \Rightarrow t = 2+3(0) = 2$$
$$1-s = -3+2t: \ 1-0 = -3+2(2) \Rightarrow 1 = 1 \ \checkmark \text{ (Note that we had to check this.)}$$

Letting $s = 0$ in the equations for the first line, or letting $t = 2$ in the equations for the second line, gives the point of intersection $(-1,2,1)$.

We will now consider planes in 3-dimensional Euclidean space.

Plane through a point, perpendicular to a vector

Let P be a plane in $\mathbb{R}^3$, and suppose it contains a point $P_0 = (x_0, y_0, z_0)$. Let $\mathbf{n} = (a, b, c)$ be a nonzero vector which is perpendicular to the plane P. Such a vector is called a **normal vector** (or just a *normal*) to the plane. Now let (x, y, z) be any point in the plane P. Then the vector $\mathbf{r} = (x - x_0, y - y_0, z - z_0)$ lies in the plane P (see Figure 1.5.6). So if $\mathbf{r} \neq \mathbf{0}$, then $\mathbf{r} \perp \mathbf{n}$ and hence $\mathbf{n} \cdot \mathbf{r} = 0$. And if $\mathbf{r} = \mathbf{0}$ then we still have $\mathbf{n} \cdot \mathbf{r} = 0$.

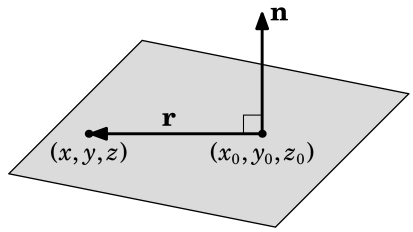


Figure 1.5.6 The plane P

Conversely, if (x, y, z) is any point in $\mathbb{R}^3$ such that $\mathbf{r} = (x - x_0, y - y_0, z - z_0) \neq \mathbf{0}$ and $\mathbf{n} \cdot \mathbf{r} = 0$, then $\mathbf{r} \perp \mathbf{n}$ and so (x, y, z) lies in P. This proves the following theorem:

Theorem 1.18. Let P be a plane in $\mathbb{R}^3$, let (x_0, y_0, z_0) be a point in P, and let $\mathbf{n} = (a, b, c)$ be a nonzero vector which is perpendicular to P. Then P consists of the points (x, y, z) satisfying the vector equation:

$$\mathbf{n} \cdot \mathbf{r} = 0 \tag{1.24}$$

where $\mathbf{r} = (x - x_0, y - y_0, z - z_0)$, or equivalently:

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0 \tag{1.25}$$

The above equation is called the **point-normal form** of the plane P.

Example 1.23. Find the equation of the plane P containing the point $(-3, 1, 3)$ and perpendicular to the vector $\mathbf{n} = (2, 4, 8)$.

Solution: By formula (1.25), the plane P consists of all points (x, y, z) such that:

$$2(x + 3) + 4(y - 1) + 8(z - 3) = 0$$

If we multiply out the terms in formula (1.25) and combine the constant terms, we get an equation of the plane in **normal form**:

$$ax + by + cz + d = 0 \tag{1.26}$$

For example, the normal form of the plane in Example 1.23 is $2x + 4y + 8z - 22 = 0$.

Plane containing three noncollinear points

In 2-dimensional and 3-dimensional space, two points determine a line. Two points do not determine a plane in $\mathbb{R}^3$. In fact, three *collinear* points (i.e. all on the same line) do not determine a plane; an infinite number of planes would contain the line on which those three points lie. However, three *noncollinear* points do determine a plane. For if Q, R and S are noncollinear points in $\mathbb{R}^3$, then $\overrightarrow{QR}$ and $\overrightarrow{QS}$ are nonzero vectors which are not parallel (by noncollinearity), and so their cross product $\overrightarrow{QR} \times \overrightarrow{QS}$ is perpendicular to both $\overrightarrow{QR}$ and $\overrightarrow{QS}$. So $\overrightarrow{QR}$ and $\overrightarrow{QS}$ (and hence Q, R and S) lie in the plane through the point Q with normal vector $\mathbf{n} = \overrightarrow{QR} \times \overrightarrow{QS}$ (see Figure 1.5.7).

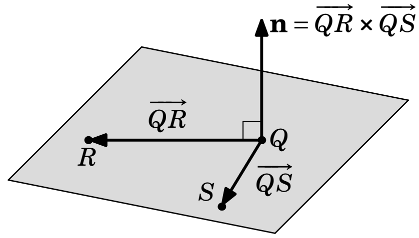


Figure 1.5.7 Noncollinear points Q, R, S

Example 1.24. Find the equation of the plane P containing the points $(2,1,3)$, $(1,-1,2)$ and $(3,2,1)$.

Solution: Let $Q = (2,1,3)$, $R = (1,-1,2)$ and $S = (3,2,1)$. Then for the vectors $\overrightarrow{QR} = (-1,-2,-1)$ and $\overrightarrow{QS} = (1,1,-2)$, the plane P has a normal vector

$$\mathbf{n} = \overrightarrow{QR} \times \overrightarrow{QS} = (-1,-2,-1) \times (1,1,-2) = (5,-3,1)$$

So using formula (1.25) with the point Q (we could also use R or S), the plane P consists of all points (x,y,z) such that:

$$5(x-2) - 3(y-1) + (z-3) = 0$$

or in normal form,

$$5x - 3y + z - 10 = 0$$

We mentioned earlier that skew lines in $\mathbb{R}^3$ lie on separate, parallel planes. So two skew lines do not determine a plane. But two (nonidentical) lines which either intersect or are parallel do determine a plane. In both cases, to find the equation of the plane that contains those two lines, simply pick from the two lines a total of three noncollinear points (i.e. one point from one line and two points from the other), then use the technique above, as in Example 1.24, to write the equation. We will leave examples of this as exercises for the reader.

Distance between a point and a plane

The distance between a point in $\mathbb{R}^3$ and a plane is the length of the line segment from that point to the plane which is perpendicular to the plane. The following theorem gives a formula for that distance.

Theorem 1.19. Let $Q = (x_0, y_0, z_0)$ be a point in $\mathbb{R}^3$, and let P be a plane with normal form $ax + by + cz + d = 0$ that does not contain Q. Then the distance D from Q to P is:

$$D = \frac{|ax_0 + by_0 + cz_0 + d|}{\sqrt{a^2 + b^2 + c^2}} \tag{1.27}$$

Proof: Let $R = (x, y, z)$ be any point in the plane P (so that $ax + by + cz + d = 0$) and let $\mathbf{r} = \overrightarrow{RQ} = (x_0 - x, y_0 - y, z_0 - z)$. Then $\mathbf{r} \neq \mathbf{0}$ since Q does not lie in P. From the normal form equation for P, we know that $\mathbf{n} = (a, b, c)$ is a normal vector for P. Now, any plane divides $\mathbb{R}^3$ into two disjoint parts. Assume that $\mathbf{n}$ points toward the side of P where the point Q is located. Place $\mathbf{n}$ so that its initial point is at R, and let θ be the angle between $\mathbf{r}$ and $\mathbf{n}$. Then $0° < \theta < 90°$, so $\cos\theta > 0$. Thus, the distance D is $\cos\theta\,\|\mathbf{r}\| = |\cos\theta|\,\|\mathbf{r}\|$ (see Figure 1.5.8).

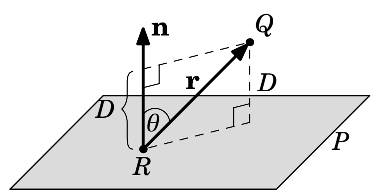


Figure 1.5.8

By Theorem 1.6 in Section 1.3, we know that $\cos\theta = \dfrac{\mathbf{n}\cdot\mathbf{r}}{\|\mathbf{n}\|\,\|\mathbf{r}\|}$, so

$$D = |\cos\theta|\,\|\mathbf{r}\| = \frac{|\mathbf{n}\cdot\mathbf{r}|}{\|\mathbf{n}\|\,\|\mathbf{r}\|}\,\|\mathbf{r}\| = \frac{|\mathbf{n}\cdot\mathbf{r}|}{\|\mathbf{n}\|} = \frac{|a(x_0 - x) + b(y_0 - y) + c(z_0 - z)|}{\sqrt{a^2 + b^2 + c^2}}$$
$$= \frac{|ax_0 + by_0 + cz_0 - (ax + by + cz)|}{\sqrt{a^2 + b^2 + c^2}} = \frac{|ax_0 + by_0 + cz_0 - (-d)|}{\sqrt{a^2 + b^2 + c^2}} = \frac{|ax_0 + by_0 + cz_0 + d|}{\sqrt{a^2 + b^2 + c^2}}$$

If $\mathbf{n}$ points away from the side of P where the point Q is located, then $90° < \theta < 180°$ and so $\cos\theta < 0$. The distance D is then $|\cos\theta|\,\|\mathbf{r}\|$, and thus repeating the same argument as above still gives the same result. **QED**

Example 1.25. Find the distance D from $(2, 4, -5)$ to the plane from Example 1.24.

Solution: Recall that the plane is given by $5x - 3y + z - 10 = 0$. So

$$D = \frac{|5(2) - 3(4) + 1(-5) - 10|}{\sqrt{5^2 + (-3)^2 + 1^2}} = \frac{|-17|}{\sqrt{35}} = \frac{17}{\sqrt{35}} \approx 2.87$$

Line of intersection of two planes

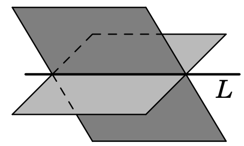


Figure 1.5.9

Note that two planes are parallel if they have normal vectors that are parallel, and the planes are perpendicular if their normal vectors are perpendicular. If two planes do intersect, they do so in a line (see Figure 1.5.9). Suppose that two planes P_1 and P_2 with normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, respectively, intersect in a line L. Since $\mathbf{n}_1 \times \mathbf{n}_2 \perp \mathbf{n}_1$, then $\mathbf{n}_1 \times \mathbf{n}_2$ is parallel to the plane P_1. Likewise, $\mathbf{n}_1 \times \mathbf{n}_2 \perp \mathbf{n}_2$ means that $\mathbf{n}_1 \times \mathbf{n}_2$ is also parallel to P_2. Thus, $\mathbf{n}_1 \times \mathbf{n}_2$ is parallel to the intersection of P_1 and P_2, i.e. $\mathbf{n}_1 \times \mathbf{n}_2$ is parallel to L. Thus, we can write L in the following vector form:

$$L : \mathbf{r} + t(\mathbf{n}_1 \times \mathbf{n}_2)\ , \text{ for } -\infty < t < \infty \tag{1.28}$$

where $\mathbf{r}$ is any vector pointing to a point belonging to both planes. To find a point in both planes, find a common solution (x, y, z) to the two normal form equations of the planes. This can often be made easier by setting one of the coordinate variables to zero, which leaves you to solve two equations in just two unknowns.

Example 1.26. Find the line of intersection L of the planes $5x - 3y + z - 10 = 0$ and $2x + 4y - z + 3 = 0$.

Solution: The plane $5x - 3y + z - 10 = 0$ has normal vector $\mathbf{n}_1 = (5, -3, 1)$ and the plane $2x + 4y - z + 3 = 0$ has normal vector $\mathbf{n}_2 = (2, 4, -1)$. Since $\mathbf{n}_1$ and $\mathbf{n}_2$ are not scalar multiples, then the two planes are not parallel and hence will intersect. A point (x, y, z) on both planes will satisfy the following system of two equations in three unknowns:

$$\begin{aligned} 5x - 3y + z - 10 &= 0 \\ 2x + 4y - z + \;\; 3 &= 0 \end{aligned}$$

Set $x = 0$ (why is that a good choice?). Then the above equations are reduced to:

$$\begin{aligned} -3y + z - 10 &= 0 \\ 4y - z + \;\; 3 &= 0 \end{aligned}$$

The second equation gives $z = 4y + 3$, substituting that into the first equation gives $y = 7$. Then $z = 31$, and so the point $(0, 7, 31)$ is on L. Since $\mathbf{n}_1 \times \mathbf{n}_2 = (-1, 7, 26)$, then L is given by:

$$\mathbf{r} + t(\mathbf{n}_1 \times \mathbf{n}_2) = (0, 7, 31) + t(-1, 7, 26), \text{ for } -\infty < t < \infty$$

or in parametric form:

$$x = -t, \quad y = 7 + 7t, \quad z = 31 + 26t, \text{ for } -\infty < t < \infty$$

Exercises

A

For Exercises 1-4, write the line L through the point P and parallel to the vector $\mathbf{v}$ in the following forms: (a) vector, (b) parametric, and (c) symmetric.

1. $P = (2,3,-2)$, $\mathbf{v} = (5,4,-3)$

2. $P = (3,-1,2)$, $\mathbf{v} = (2,8,1)$

3. $P = (2,1,3)$, $\mathbf{v} = (1,0,1)$

4. $P = (0,0,0)$, $\mathbf{v} = (7,2,-10)$

For Exercises 5-6, write the line L through the points P_1 and P_2 in parametric form.

5. $P_1 = (1,-2,-3)$, $P_2 = (3,5,5)$

6. $P_1 = (4,1,5)$, $P_2 = (-2,1,3)$

For Exercises 7-8, find the distance d from the point P to the line L.

7. $P = (1,-1,-1)$, $L: x = -2-2t,\ y = 4t,\ z = 7+t$

8. $P = (0,0,0)$, $L: x = 3+2t,\ y = 4+3t,\ z = 5+4t$

For Exercises 9-10, find the point of intersection (if any) of the given lines.

9. $x = 7+3s,\ y = -4-3s,\ z = -7-5s$ and $x = 1+6t,\ y = 2+t,\ z = 3-2t$

10. $\dfrac{x-6}{4} = y+3 = z$ and $\dfrac{x-11}{3} = \dfrac{y-14}{-6} = \dfrac{z+9}{2}$

For Exercises 11-12, write the normal form of the plane P containing the point Q and perpendicular to the vector $\mathbf{n}$.

11. $Q = (5,1,-2)$, $\mathbf{n} = (4,-4,3)$

12. $Q = (6,-2,0)$, $\mathbf{n} = (2,6,4)$

For Exercises 13-14, write the normal form of the plane containing the given points.

13. $(1,0,3)$, $(1,2,-1)$, $(6,1,6)$

14. $(-3,1,-3)$, $(4,-4,3)$, $(0,0,1)$

15. Write the normal form of the plane containing the lines from Exercise 9.

16. Write the normal form of the plane containing the lines from Exercise 10.

For Exercises 17-18, find the distance D from the point Q to the plane P.

17. $Q = (4,1,2)$, $P: 3x-y-5z+8 = 0$

18. $Q = (0,2,0)$, $P: -5x+2y-7z+1 = 0$

For Exercises 19-20, find the line of intersection (if any) of the given planes.

19. $x+3y+2z-6 = 0$, $2x-y+z+2 = 0$

20. $3x+y-5z = 0$, $x+2y+z+4 = 0$

B

21. Find the point(s) of intersection (if any) of the line $\dfrac{x-6}{4} = y+3 = z$ with the plane $x+3y+2z-6 = 0$. (*Hint: Put the equations of the line into the equation of the plane.*)

1.6 Surfaces

In the previous section we discussed planes in Euclidean space. A plane is an example of a *surface*, which we will define informally[8] as the solution set of the equation $F(x,y,z)=0$ in $\mathbb{R}^3$, for some real-valued function F. For example, a plane given by $ax+by+cz+d=0$ is the solution set of $F(x,y,z)=0$ for the function $F(x,y,z)=ax+by+cz+d$. Surfaces are 2-dimensional. The plane is the simplest surface, since it is "flat". In this section we will look at some surfaces that are more complex, the most important of which are the sphere and the cylinder.

Definition 1.9. A **sphere** S is the set of all points (x,y,z) in $\mathbb{R}^3$ which are a fixed distance r (called the **radius**) from a fixed point $P_0=(x_0,y_0,z_0)$ (called the **center** of the sphere):

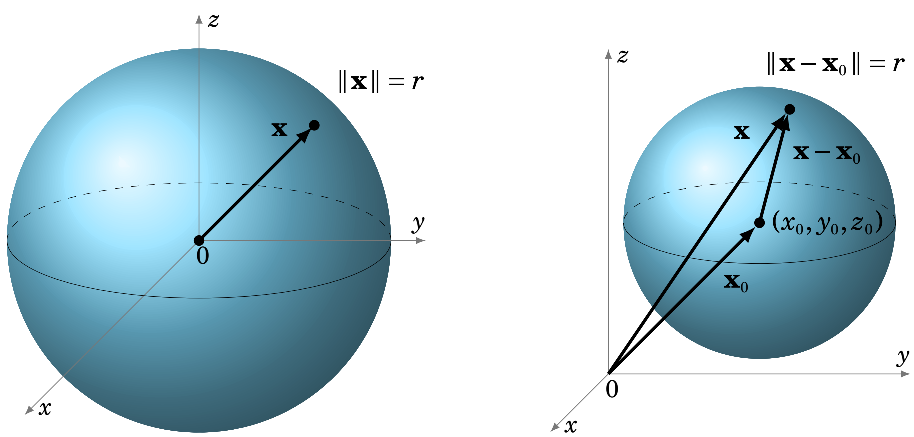
$$S=\{(x,y,z):(x-x_0)^2+(y-y_0)^2+(z-z_0)^2=r^2\} \tag{1.29}$$

Using vector notation, this can be written in the equivalent form:

$$S=\{\mathbf{x}:\|\mathbf{x}-\mathbf{x}_0\|=r\} \tag{1.30}$$

where $\mathbf{x}=(x,y,z)$ and $\mathbf{x}_0=(x_0,y_0,z_0)$ are vectors.

Figure 1.6.1 illustrates the vectorial approach to spheres.



(a) radius r, center $(0,0,0)$ **(b)** radius r, center (x_0,y_0,z_0)

Figure 1.6.1 Spheres in $\mathbb{R}^3$

Note in Figure 1.6.1(a) that the intersection of the sphere with the xy-plane is a circle of radius r (i.e. a *great circle*, given by $x^2+y^2=r^2$ as a subset of $\mathbb{R}^2$). Similarly for the intersections with the xz-plane and the yz-plane. In general, a plane intersects a sphere either at a single point or in a circle.

[8]See O'NEILL for a deeper and more rigorous discussion of surfaces.

Example 1.27. Find the intersection of the sphere $x^2 + y^2 + z^2 = 169$ with the plane $z = 12$.

Solution: The sphere is centered at the origin and has radius $13 = \sqrt{169}$, so it does intersect the plane $z = 12$. Putting $z = 12$ into the equation of the sphere gives

$$\begin{aligned} x^2 + y^2 + 12^2 &= 169 \\ x^2 + y^2 &= 169 - 144 = 25 = 5^2 \end{aligned}$$

which is a circle of radius 5 centered at $(0,0,12)$, parallel to the xy-plane (see Figure 1.6.2).

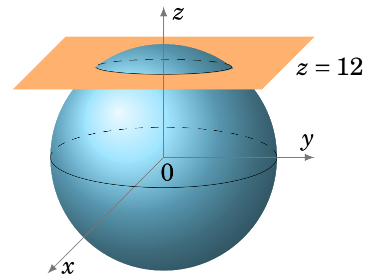


Figure 1.6.2

If the equation in formula (1.29) is multiplied out, we get an equation of the form:

$$x^2 + y^2 + z^2 + ax + by + cz + d = 0 \tag{1.31}$$

for some constants a, b, c and d. Conversely, an equation of this form *may* describe a sphere, which can be determined by completing the square for the x, y and z variables.

Example 1.28. Is $2x^2 + 2y^2 + 2z^2 - 8x + 4y - 16z + 10 = 0$ the equation of a sphere?

Solution: Dividing both sides of the equation by 2 gives

$$\begin{aligned} x^2 + y^2 + z^2 - 4x + 2y - 8z + 5 &= 0 \\ (x^2 - 4x + 4) + (y^2 + 2y + 1) + (z^2 - 8z + 16) + 5 - 4 - 1 - 16 &= 0 \\ (x-2)^2 + (y+1)^2 + (z-4)^2 &= 16 \end{aligned}$$

which is a sphere of radius 4 centered at $(2,-1,4)$.

Example 1.29. Find the points(s) of intersection (if any) of the sphere from Example 1.28 and the line $x = 3 + t$, $y = 1 + 2t$, $z = 3 - t$.

Solution: Put the equations of the line into the equation of the sphere, which was $(x-2)^2 + (y+1)^2 + (z-4)^2 = 16$, and solve for t:

$$\begin{aligned} (3 + t - 2)^2 + (1 + 2t + 1)^2 + (3 - t - 4)^2 &= 16 \\ (t+1)^2 + (2t+2)^2 + (-t-1)^2 &= 16 \\ 6t^2 + 12t - 10 &= 0 \end{aligned}$$

The quadratic formula gives the solutions $t = -1 \pm \frac{4}{\sqrt{6}}$. Putting those two values into the equations of the line gives the following two points of intersection:

$$\left(2 + \frac{4}{\sqrt{6}}, -1 + \frac{8}{\sqrt{6}}, 4 - \frac{4}{\sqrt{6}}\right) \quad \text{and} \quad \left(2 - \frac{4}{\sqrt{6}}, -1 - \frac{8}{\sqrt{6}}, 4 + \frac{4}{\sqrt{6}}\right)$$

If two spheres intersect, they do so either at a single point or in a circle.

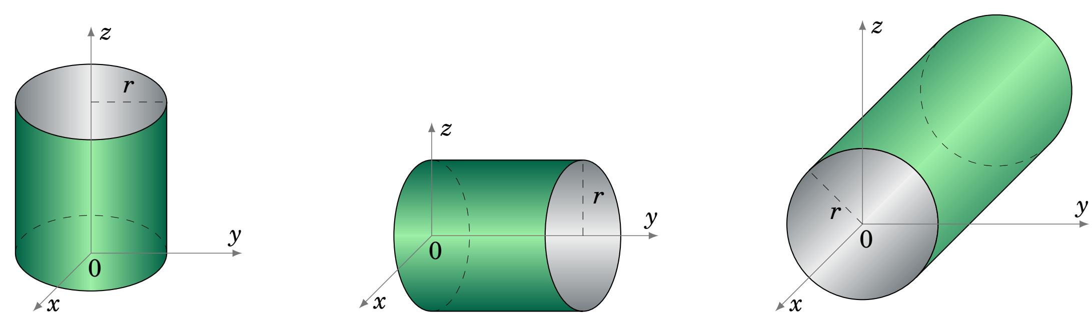
Example 1.30. Find the intersection (if any) of the spheres $x^2+y^2+z^2=25$ and $x^2+y^2+(z-2)^2=16$.

Solution: For any point (x,y,z) on both spheres, we see that

$$\begin{aligned}
x^2+y^2+z^2=25 \quad &\Rightarrow \quad x^2+y^2=25-z^2, \text{ and}\\
x^2+y^2+(z-2)^2=16 \quad &\Rightarrow \quad x^2+y^2=16-(z-2)^2, \text{ so}\\
16-(z-2)^2=25-z^2 \quad &\Rightarrow \quad 4z-4=9 \quad \Rightarrow \quad z=13/4\\
&\Rightarrow \quad x^2+y^2=25-(13/4)^2=231/16
\end{aligned}$$

$\therefore$ The intersection is the circle $x^2+y^2=\frac{231}{16}$ of radius $\frac{\sqrt{231}}{4}\approx 3.8$ centered at $(0,0,\frac{13}{4})$.

The cylinders that we will consider are *right circular cylinders*. These are cylinders obtained by moving a line L along a circle C in $\mathbb{R}^3$ in a way so that L is always perpendicular to the plane containing C. We will only consider the cases where the plane containing C is parallel to one of the three coordinate planes (see Figure 1.6.3).



(a) $x^2+y^2=r^2$, any z **(b)** $x^2+z^2=r^2$, any y **(c)** $y^2+z^2=r^2$, any x

Figure 1.6.3 Cylinders in $\mathbb{R}^3$

For example, the equation of a cylinder whose base circle C lies in the xy-plane and is centered at $(a,b,0)$ and has radius r is

$$(x-a)^2+(y-b)^2=r^2, \tag{1.32}$$

where the value of the z coordinate is unrestricted. Similar equations can be written when the base circle lies in one of the other coordinate planes. A plane intersects a right circular cylinder in a circle, ellipse, or one or two lines, depending on whether that plane is parallel, oblique[9], or perpendicular, respectively, to the plane containing C. The intersection of a surface with a plane is called the **trace** of the surface.

[9]i.e. at an angle strictly between 0° and 90°.

The equations of spheres and cylinders are examples of *second-degree equations* in $\mathbb{R}^3$, i.e. equations of the form

$$Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz + Gx + Hy + Iz + J = 0 \tag{1.33}$$

for some constants $A, B, \ldots, J$. If the above equation is not that of a sphere, cylinder, plane, line or point, then the resulting surface is called a **quadric surface**.

One type of quadric surface is the **ellipsoid**, given by an equation of the form:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \tag{1.34}$$

In the case where $a = b = c$, this is just a sphere. In general, an ellipsoid is egg-shaped (think of an ellipse rotated around its major axis). Its traces in the coordinate planes are ellipses.

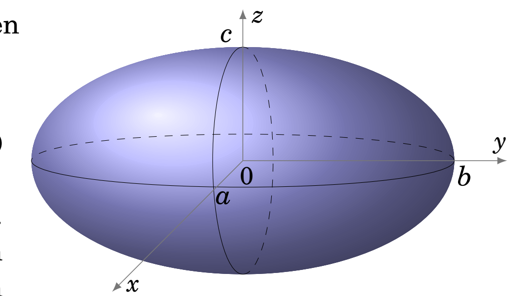


Figure 1.6.4 Ellipsoid

Two other types of quadric surfaces are the **hyperboloid of one sheet**, given by an equation of the form:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \tag{1.35}$$

and the **hyperboloid of two sheets**, whose equation has the form:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \tag{1.36}$$

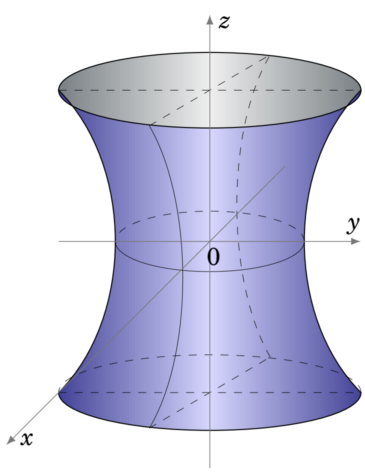


Figure 1.6.5 Hyperboloid of one sheet

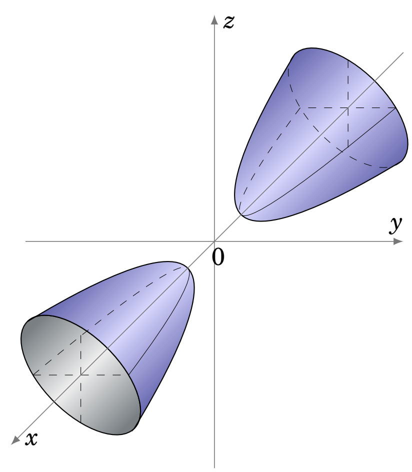


Figure 1.6.6 Hyperboloid of two sheets

For the hyperboloid of one sheet, the trace in any plane parallel to the xy-plane is an ellipse. The traces in the planes parallel to the xz- or yz-planes are hyperbolas (see Figure 1.6.5), except for the special cases $x = \pm a$ and $y = \pm b$; in those planes the traces are pairs of intersecting lines (see Exercise 8).

For the hyperboloid of two sheets, the trace in any plane parallel to the xy- or xz-plane is a hyperbola (see Figure 1.6.6). There is no trace in the yz-plane. In any plane parallel to the yz-plane for which $|x| > |a|$, the trace is an ellipse.

The **elliptic paraboloid** is another type of quadric surface, whose equation has the form:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z}{c} \tag{1.37}$$

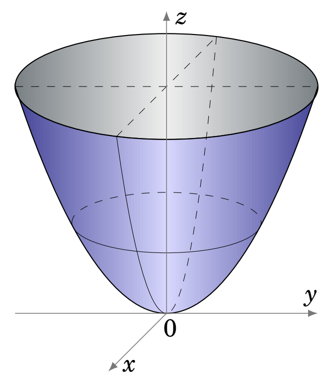


Figure 1.6.7 Paraboloid

The traces in planes parallel to the xy-plane are ellipses, though in the xy-plane itself the trace is a single point. The traces in planes parallel to the xz- or yz-planes are parabolas. Figure 1.6.7 shows the case where $c > 0$. When $c < 0$ the surface is turned downward. In the case where $a = b$, the surface is called a *paraboloid of revolution*, which is often used as a reflecting surface, e.g. in vehicle headlights.[10]

A more complicated quadric surface is the **hyperbolic paraboloid**, given by:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = \frac{z}{c} \tag{1.38}$$

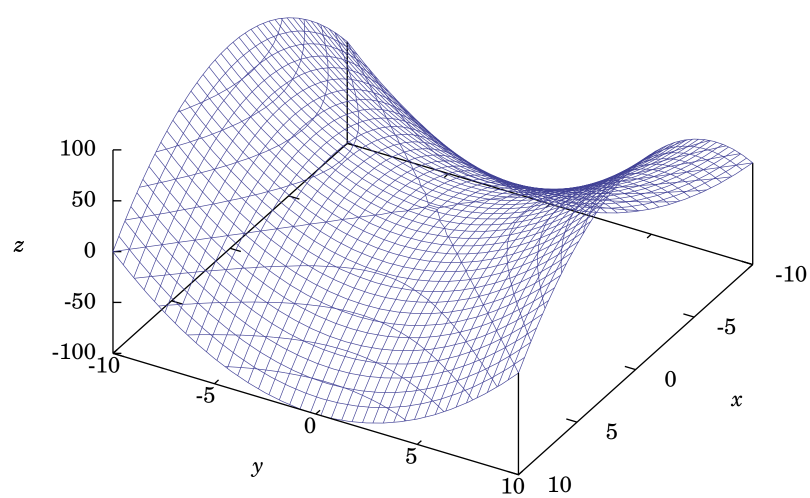


Figure 1.6.8 Hyperbolic paraboloid

[10] For a discussion of this see pp. 157-158 in HECHT.

The hyperbolic paraboloid can be tricky to draw; using graphing software on a computer can make it easier. For example, Figure 1.6.8 was created using the free Gnuplot package (see Appendix C). It shows the graph of the hyperbolic paraboloid $z = y^2 - x^2$, which is the special case where $a = b = 1$ and $c = -1$ in equation (1.38). The mesh lines on the surface are the traces in planes parallel to the coordinate planes. So we see that the traces in planes parallel to the xz-plane are parabolas pointing upward, while the traces in planes parallel to the yz-plane are parabolas pointing downward. Also, notice that the traces in planes parallel to the xy-plane are hyperbolas, though in the xy-plane itself the trace is a pair of intersecting lines through the origin. This is true in general when $c < 0$ in equation (1.38). When $c > 0$, the surface would be similar to that in Figure 1.6.8, only rotated 90° around the z-axis and the nature of the traces in planes parallel to the xz- or yz-planes would be reversed.

The last type of quadric surface that we will consider is the **elliptic cone**, which has an equation of the form:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0 \tag{1.39}$$

The traces in planes parallel to the xy-plane are ellipses, except in the xy-plane itself where the trace is a single point. The traces in planes parallel to the xz- or yz-planes are hyperbolas, except in the xz- and yz-planes themselves where the traces are pairs of intersecting lines.

Notice that every point on the elliptic cone is on a line which lies entirely on the surface; in Figure 1.6.9 these lines all go through the origin. This makes the elliptic cone an example of a *ruled surface*. The cylinder is also a ruled surface.

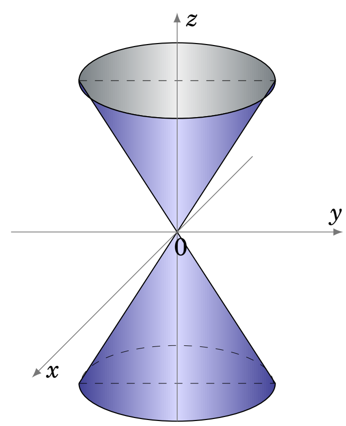


Figure 1.6.9 Elliptic cone

What may not be as obvious is that both the hyperboloid of one sheet and the hyperbolic paraboloid are ruled surfaces. In fact, on both surfaces there are *two* lines through each point on the surface (see Exercises 11-12). Such surfaces are called *doubly ruled surfaces*, and the pairs of lines are called a *regulus*.

It is clear that for each of the six types of quadric surfaces that we discussed, the surface can be translated away from the origin (e.g. by replacing x^2 by $(x - x_0)^2$ in its equation). It can be proved[11] that *every* quadric surface can be translated and/or rotated so that its equation matches one of the six types that we described. For example, $z = 2xy$ is a case of equation (1.33) with "mixed" variables, e.g. with $D \neq 0$ so that we get an xy term. This equation does not match any of the types we considered. However, by rotating the x- and y-axes by 45° in the xy-plane by means of the coordinate transformation $x = (x' - y')/\sqrt{2}$, $y = (x' + y')/\sqrt{2}$, $z = z'$, then $z = 2xy$ becomes the hyperbolic paraboloid $z' = (x')^2 - (y')^2$ in the (x', y', z') coordinate system. That is, $z = 2xy$ is a hyperbolic paraboloid as in equation (1.38), but rotated 45° in the xy-plane.

[11] See Ch. 7 in POGORELOV.

Exercises

A

For Exercises 1-4, determine if the given equation describes a sphere. If so, find its radius and center.

1. $x^2 + y^2 + z^2 - 4x - 6y - 10z + 37 = 0$

2. $x^2 + y^2 + z^2 + 2x - 2y - 8z + 19 = 0$

3. $2x^2 + 2y^2 + 2z^2 + 4x + 4y + 4z - 44 = 0$

4. $x^2 + y^2 - z^2 + 12x + 2y - 4z + 32 = 0$

5. Find the point(s) of intersection of the sphere $(x-3)^2 + (y+1)^2 + (z-3)^2 = 9$ and the line $x = -1 + 2t$, $y = -2 - 3t$, $z = 3 + t$.

B

6. Find the intersection of the spheres $x^2 + y^2 + z^2 = 9$ and $(x-4)^2 + (y+2)^2 + (z-4)^2 = 9$.

7. Find the intersection of the sphere $x^2 + y^2 + z^2 = 9$ and the cylinder $x^2 + y^2 = 4$.

8. Find the trace of the hyperboloid of one sheet $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$ in the plane $x = a$, and the trace in the plane $y = b$.

9. Find the trace of the hyperbolic paraboloid $\frac{x^2}{a^2} - \frac{y^2}{b^2} = \frac{z}{c}$ in the xy-plane.

C

10. It can be shown that any four *noncoplanar* points (i.e. points that do not lie in the same plane) determine a sphere.[12] Find the equation of the sphere that passes through the points $(0,0,0)$, $(0,0,2)$, $(1,-4,3)$ and $(0,-1,3)$. *(Hint: Equation (1.31))*

11. Show that the hyperboloid of one sheet is a doubly ruled surface, i.e. each point on the surface is on two lines lying entirely on the surface. *(Hint: Write equation (1.35) as $\frac{x^2}{a^2} - \frac{z^2}{c^2} = 1 - \frac{y^2}{b^2}$, factor each side. Recall that two planes intersect in a line.)*

12. Show that the hyperbolic paraboloid is a doubly ruled surface. *(Hint: Exercise 11)*

13. Let S be the sphere with radius 1 centered at $(0,0,1)$, and let S^* be S without the "north pole" point $(0,0,2)$. Let (a,b,c) be an arbitrary point on S^*. Then the line passing through $(0,0,2)$ and (a,b,c) intersects the xy-plane at some point $(x,y,0)$, as in Figure 1.6.10. Find this point $(x,y,0)$ in terms of a, b and c.
(Note: Every point in the xy-plane can be matched with a point on S^*, and vice versa, in this manner. This method is called *stereographic projection*, which essentially identifies all of $\mathbb{R}^2$ with a "punctured" sphere.)

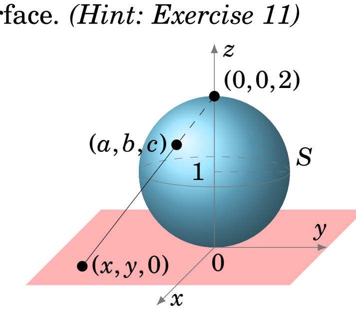


Figure 1.6.10

[12]See WELCHONS and KRICKENBERGER, p. 160, for a proof.

1.7 Curvilinear Coordinates

The Cartesian coordinates of a point (x, y, z) are determined by following straight paths starting from the origin: first along the x-axis, then parallel to the y-axis, then parallel to the z-axis, as in Figure 1.7.1. In *curvilinear coordinate systems*, these paths can be curved. The two types of curvilinear coordinates which we will consider are cylindrical and spherical coordinates. Instead of referencing a point in terms of sides of a rectangular parallelepiped, as with Cartesian coordinates, we will think of the point as lying on a cylinder or sphere. Cylindrical coordinates are often used when there is symmetry around the z-axis; spherical coordinates are useful when there is symmetry about the origin.

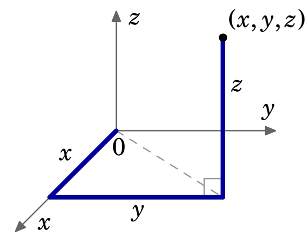


Figure 1.7.1

Let $P = (x, y, z)$ be a point in Cartesian coordinates in $\mathbb{R}^3$, and let $P_0 = (x, y, 0)$ be the projection of P upon the xy-plane. Treating (x, y) as a point in $\mathbb{R}^2$, let (r, θ) be its polar coordinates (see Figure 1.7.2). Let ρ be the length of the line segment from the origin to P, and let ϕ be the angle between that line segment and the positive z-axis (see Figure 1.7.3). ϕ is called the *zenith angle*. Then the **cylindrical coordinates** (r, θ, z) and the **spherical coordinates** (ρ, θ, ϕ) of $P(x, y, z)$ are defined as follows:[13]

Cylindrical coordinates (r, θ, z):

$$x = r\cos\theta \qquad r = \sqrt{x^2 + y^2}$$
$$y = r\sin\theta \qquad \theta = \tan^{-1}\left(\tfrac{y}{x}\right)$$
$$z = z \qquad z = z$$

where $0 \le \theta \le \pi$ if $y \ge 0$ and $\pi < \theta < 2\pi$ if $y < 0$

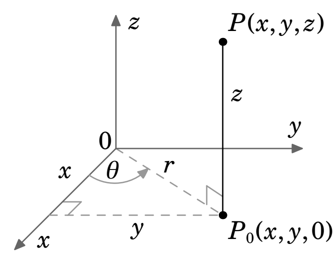


Figure 1.7.2
Cylindrical coordinates

Spherical coordinates (ρ, θ, ϕ):

$$x = \rho\sin\phi\cos\theta \qquad \rho = \sqrt{x^2 + y^2 + z^2}$$
$$y = \rho\sin\phi\sin\theta \qquad \theta = \tan^{-1}\left(\tfrac{y}{x}\right)$$
$$z = \rho\cos\phi \qquad \phi = \cos^{-1}\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right)$$

where $0 \le \theta \le \pi$ if $y \ge 0$ and $\pi < \theta < 2\pi$ if $y < 0$

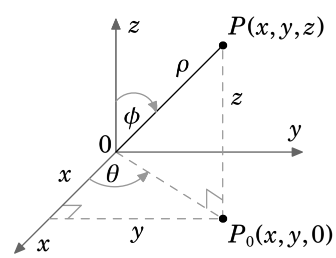


Figure 1.7.3
Spherical coordinates

Both θ and ϕ are measured in radians. Note that $r \ge 0$, $0 \le \theta < 2\pi$, $\rho \ge 0$ and $0 \le \phi \le \pi$. Also, θ is undefined when $(x, y) = (0, 0)$, and ϕ is undefined when $(x, y, z) = (0, 0, 0)$.

[13] This "standard" definition of spherical coordinates used by mathematicians results in a left-handed system. For this reason, physicists usually switch the definitions of θ and ϕ to make (ρ, θ, ϕ) a right-handed system.

Example 1.31. Convert the point $(-2,-2,1)$ from Cartesian coordinates to (a) cylindrical and (b) spherical coordinates.

Solution: (a) $r = \sqrt{(-2)^2+(-2)^2} = 2\sqrt{2}$, $\theta = \tan^{-1}\left(\frac{-2}{-2}\right) = \tan^{-1}(1) = \frac{5\pi}{4}$, since $y = -2 < 0$.
$\therefore\ (r,\theta,z) = \left(2\sqrt{2}, \frac{5\pi}{4}, 1\right)$

(b) $\rho = \sqrt{(-2)^2+(-2)^2+1^2} = \sqrt{9} = 3$, $\phi = \cos^{-1}\left(\frac{1}{3}\right) \approx 1.23$ radians.
$\therefore\ (\rho,\theta,\phi) = \left(3, \frac{5\pi}{4}, 1.23\right)$

For cylindrical coordinates (r,θ,z), and constants r_0, θ_0 and z_0, we see from Figure 1.7.4 that the surface $r = r_0$ is a cylinder of radius r_0 centered along the z-axis, the surface $\theta = \theta_0$ is a half-plane emanating from the z-axis, and the surface $z = z_0$ is a plane parallel to the xy-plane.

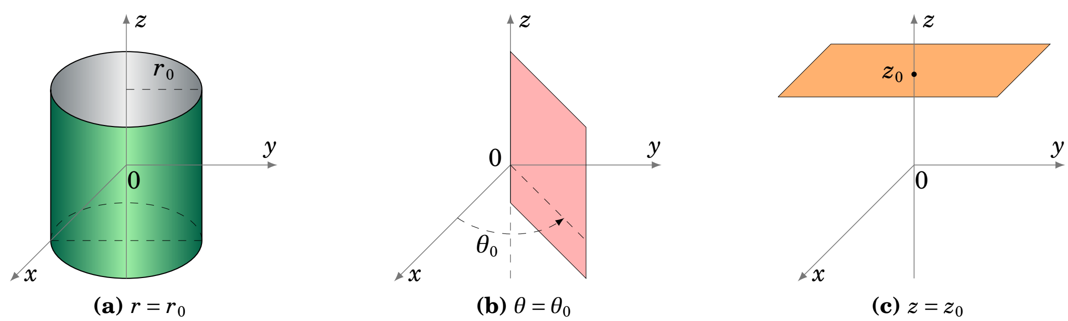


Figure 1.7.4 Cylindrical coordinate surfaces

For spherical coordinates (ρ,θ,ϕ), and constants ρ_0, θ_0 and ϕ_0, we see from Figure 1.7.5 that the surface $\rho = \rho_0$ is a sphere of radius ρ_0 centered at the origin, the surface $\theta = \theta_0$ is a half-plane emanating from the z-axis, and the surface $\phi = \phi_0$ is a circular cone whose vertex is at the origin.

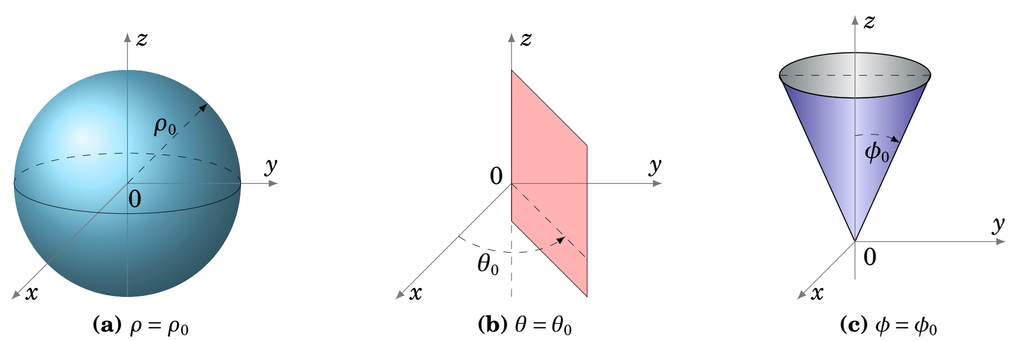


Figure 1.7.5 Spherical coordinate surfaces

Figures 1.7.4(a) and 1.7.5(a) show how these coordinate systems got their names.

Sometimes the equation of a surface in Cartesian coordinates can be transformed into a simpler equation in some other coordinate system, as in the following example.

Example 1.32. Write the equation of the cylinder $x^2 + y^2 = 4$ in cylindrical coordinates.

Solution: Since $r = \sqrt{x^2 + y^2}$, then the equation in cylindrical coordinates is $r = 2$.

Using spherical coordinates to write the equation of a sphere does not necessarily make the equation simpler, if the sphere is not centered at the origin.

Example 1.33. Write the equation $(x-2)^2 + (y-1)^2 + z^2 = 9$ in spherical coordinates.

Solution: Multiplying the equation out gives

$$x^2 + y^2 + z^2 - 4x - 2y + 5 = 9 \text{ , so we get}$$
$$\rho^2 - 4\rho \sin\phi \cos\theta - 2\rho \sin\phi \sin\theta - 4 = 0 \text{ , or}$$
$$\rho^2 - 2\sin\phi (2\cos\theta - \sin\theta)\, \rho - 4 = 0$$

after combining terms. Note that this actually makes it more difficult to figure out what the surface is, as opposed to the Cartesian equation where you could immediately identify the surface as a sphere of radius 3 centered at $(2,1,0)$.

Example 1.34. Describe the surface given by $\theta = z$ in cylindrical coordinates.

Solution: This surface is called a *helicoid*. As the (vertical) z coordinate increases, so does the angle θ, while the radius r is unrestricted. So this sweeps out a (ruled!) surface shaped like a spiral staircase, where the spiral has an infinite radius. Figure 1.7.6 shows a section of this surface restricted to $0 \le z \le 4\pi$ and $0 \le r \le 2$.

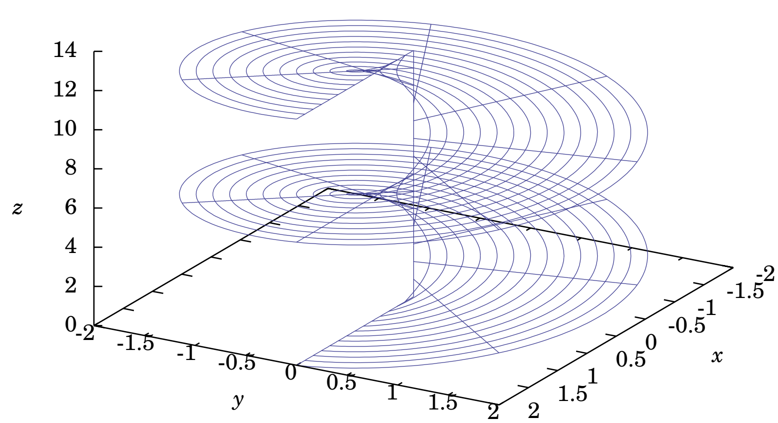


Figure 1.7.6 Helicoid $\theta = z$

Exercises

A

For Exercises 1-4, find the (a) cylindrical and (b) spherical coordinates of the point whose Cartesian coordinates are given.

1. $(2, 2\sqrt{3}, -1)$ **2.** $(-5, 5, 6)$ **3.** $(\sqrt{21}, -\sqrt{7}, 0)$ **4.** $(0, \sqrt{2}, 2)$

For Exercises 5-7, write the given equation in (a) cylindrical and (b) spherical coordinates.

5. $x^2 + y^2 + z^2 = 25$ **6.** $x^2 + y^2 = 2y$ **7.** $x^2 + y^2 + 9z^2 = 36$

B

8. Describe the intersection of the surfaces whose equations in spherical coordinates are $\theta = \frac{\pi}{2}$ and $\phi = \frac{\pi}{4}$.

9. Show that for $a \neq 0$, the equation $\rho = 2a \sin\phi \cos\theta$ in spherical coordinates describes a sphere centered at $(a, 0, 0)$ with radius $|a|$.

C

10. Let $P = (a, \theta, \phi)$ be a point in spherical coordinates, with $a > 0$ and $0 < \phi < \pi$. Then P lies on the sphere $\rho = a$. Since $0 < \phi < \pi$, the line segment from the origin to P can be extended to intersect the cylinder given by $r = a$ (in cylindrical coordinates). Find the cylindrical coordinates of that point of intersection.

11. Let P_1 and P_2 be points whose spherical coordinates are $(\rho_1, \theta_1, \phi_1)$ and $(\rho_2, \theta_2, \phi_2)$, respectively. Let $\mathbf{v}_1$ be the vector from the origin to P_1, and let $\mathbf{v}_2$ be the vector from the origin to P_2. For the angle γ between $\mathbf{v}_1$ and $\mathbf{v}_2$, show that

$$\cos\gamma = \cos\phi_1 \cos\phi_2 + \sin\phi_1 \sin\phi_2 \cos(\theta_2 - \theta_1).$$

This formula is used in electrodynamics to prove the addition theorem for spherical harmonics, which provides a general expression for the electrostatic potential at a point due to a unit charge. See pp. 100-102 in JACKSON.

12. Show that the distance d between the points P_1 and P_2 with cylindrical coordinates (r_1, θ_1, z_1) and (r_2, θ_2, z_2), respectively, is

$$d = \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_2 - \theta_1) + (z_2 - z_1)^2}.$$

13. Show that the distance d between the points P_1 and P_2 with spherical coordinates $(\rho_1, \theta_1, \phi_1)$ and $(\rho_2, \theta_2, \phi_2)$, respectively, is

$$d = \sqrt{\rho_1^2 + \rho_2^2 - 2\rho_1 \rho_2 [\sin\phi_1 \sin\phi_2 \cos(\theta_2 - \theta_1) + \cos\phi_1 \cos\phi_2]}.$$

1.8 Vector-Valued Functions

Now that we are familiar with vectors and their operations, we can begin discussing functions whose values are vectors.

Definition 1.10. A **vector-valued function of a real variable** is a rule that associates a vector $\mathbf{f}(t)$ with a real number t, where t is in some subset D of $\mathbb{R}^1$ (called the **domain** of $\mathbf{f}$). We write $\mathbf{f}: D \to \mathbb{R}^3$ to denote that $\mathbf{f}$ is a mapping of D into $\mathbb{R}^3$.

For example, $\mathbf{f}(t) = t\mathbf{i} + t^2\mathbf{j} + t^3\mathbf{k}$ is a vector-valued function in $\mathbb{R}^3$, defined for all real numbers t. We would write $\mathbf{f}: \mathbb{R} \to \mathbb{R}^3$. At $t = 1$ the value of the function is the vector $\mathbf{i}+\mathbf{j}+\mathbf{k}$, which in Cartesian coordinates has the terminal point $(1,1,1)$.

A vector-valued function of a real variable can be written in component form as

$$\mathbf{f}(t) = f_1(t)\mathbf{i} + f_2(t)\mathbf{j} + f_3(t)\mathbf{k}$$

or in the form

$$\mathbf{f}(t) = (f_1(t), f_2(t), f_3(t))$$

for some real-valued functions $f_1(t)$, $f_2(t)$, $f_3(t)$, called the *component functions* of $\mathbf{f}$. The first form is often used when emphasizing that $\mathbf{f}(t)$ is a vector, and the second form is useful when considering just the terminal points of the vectors. By identifying vectors with their terminal points, a curve in space can be written as a vector-valued function.

Example 1.35. Define $\mathbf{f}: \mathbb{R} \to \mathbb{R}^3$ by $\mathbf{f}(t) = (\cos t, \sin t, t)$.
This is the equation of a *helix* (see Figure 1.8.1). As the value of t increases, the terminal points of $\mathbf{f}(t)$ trace out a curve spiraling upward. For each t, the x- and y-coordinates of $\mathbf{f}(t)$ are $x = \cos t$ and $y = \sin t$, so

$$x^2 + y^2 = \cos^2 t + \sin^2 t = 1.$$

Thus, the curve lies on the surface of the right circular cylinder $x^2 + y^2 = 1$.

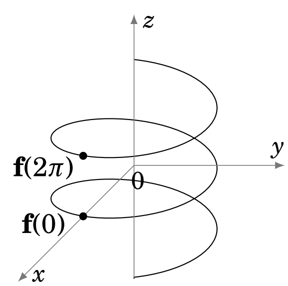


Figure 1.8.1

It may help to think of vector-valued functions of a real variable in $\mathbb{R}^3$ as a generalization of the parametric functions in $\mathbb{R}^2$ which you learned about in single-variable calculus. Much of the theory of real-valued functions of a single real variable can be applied to vector-valued functions of a real variable. Since each of the three component functions are real-valued, it will sometimes be the case that results from single-variable calculus can simply be applied to each of the component functions to yield a similar result for the vector-valued function. However, there are times when such generalizations do not hold (see Exercise 13). The concept of a limit, though, can be extended naturally to vector-valued functions, as in the following definition.

Definition 1.11. Let $\mathbf{f}(t)$ be a vector-valued function, let a be a real number and let $\mathbf{c}$ be a vector. Then we say that the **limit** of $\mathbf{f}(t)$ as t approaches a equals $\mathbf{c}$, written as $\lim_{t \to a} \mathbf{f}(t) = \mathbf{c}$, if $\lim_{t \to a} \|\mathbf{f}(t) - \mathbf{c}\| = 0$. If $\mathbf{f}(t) = (f_1(t), f_2(t), f_3(t))$, then

$$\lim_{t \to a} \mathbf{f}(t) = \left(\lim_{t \to a} f_1(t), \lim_{t \to a} f_2(t), \lim_{t \to a} f_3(t)\right)$$

provided that all three limits on the right side exist.

The above definition shows that continuity and the derivative of vector-valued functions can also be defined in terms of its component functions.

Definition 1.12. Let $\mathbf{f}(t) = (f_1(t), f_2(t), f_3(t))$ be a vector-valued function, and let a be a real number in its domain. Then $\mathbf{f}(t)$ is **continuous** at a if $\lim_{t \to a} \mathbf{f}(t) = \mathbf{f}(a)$. Equivalently, $\mathbf{f}(t)$ is continuous at a if and only if $f_1(t)$, $f_2(t)$, and $f_3(t)$ are continuous at a.
The **derivative** of $\mathbf{f}(t)$ at a, denoted by $\mathbf{f}'(a)$ or $\dfrac{d\mathbf{f}}{dt}(a)$, is the limit

$$\mathbf{f}'(a) = \lim_{h \to 0} \frac{\mathbf{f}(a+h) - \mathbf{f}(a)}{h}$$

if that limit exists. Equivalently, $\mathbf{f}'(a) = (f_1{}'(a), f_2{}'(a), f_3{}'(a))$, if the component derivatives exist. We say that $\mathbf{f}(t)$ is **differentiable** at a if $\mathbf{f}'(a)$ exists.

Recall that the derivative of a real-valued function of a single variable is a real number, representing the slope of the tangent line to the graph of the function at a point. Similarly, the derivative of a vector-valued function is a **tangent vector** to the curve in space which the function represents, and it lies on the *tangent line* to the curve (see Figure 1.8.2).

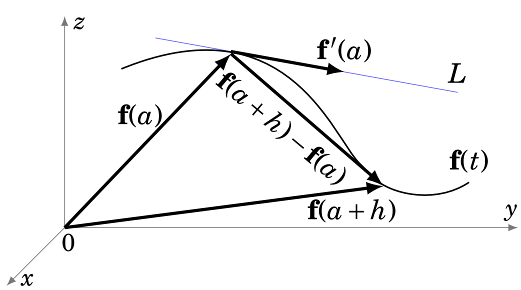


Figure 1.8.2 Tangent vector $\mathbf{f}'(a)$ and tangent line $L = \mathbf{f}(a) + s\mathbf{f}'(a)$

Example 1.36. Let $\mathbf{f}(t) = (\cos t, \sin t, t)$. Then $\mathbf{f}'(t) = (-\sin t, \cos t, 1)$ for all t. The tangent line L to the curve at $\mathbf{f}(2\pi) = (1, 0, 2\pi)$ is $L = \mathbf{f}(2\pi) + s\,\mathbf{f}'(2\pi) = (1, 0, 2\pi) + s(0, 1, 1)$, or in parametric form: $x = 1$, $y = s$, $z = 2\pi + s$ for $-\infty < s < \infty$.

A **scalar function** is a real-valued function. Note that if $u(t)$ is a scalar function and $\mathbf{f}(t)$ is a vector-valued function, then their product, defined by $(u\,\mathbf{f})(t) = u(t)\mathbf{f}(t)$ for all t, is a vector-valued function (since the product of a scalar with a vector is a vector).

The basic properties of derivatives of vector-valued functions are summarized in the following theorem.

Theorem 1.20. Let $\mathbf{f}(t)$ and $\mathbf{g}(t)$ be differentiable vector-valued functions, let $u(t)$ be a differentiable scalar function, let k be a scalar, and let $\mathbf{c}$ be a constant vector. Then

(a) $\dfrac{d}{dt}(\mathbf{c}) = \mathbf{0}$

(b) $\dfrac{d}{dt}(k\mathbf{f}) = k\dfrac{d\mathbf{f}}{dt}$

(c) $\dfrac{d}{dt}(\mathbf{f}+\mathbf{g}) = \dfrac{d\mathbf{f}}{dt} + \dfrac{d\mathbf{g}}{dt}$

(d) $\dfrac{d}{dt}(\mathbf{f}-\mathbf{g}) = \dfrac{d\mathbf{f}}{dt} - \dfrac{d\mathbf{g}}{dt}$

(e) $\dfrac{d}{dt}(u\,\mathbf{f}) = \dfrac{du}{dt}\mathbf{f} + u\dfrac{d\mathbf{f}}{dt}$

(f) $\dfrac{d}{dt}(\mathbf{f}\cdot\mathbf{g}) = \dfrac{d\mathbf{f}}{dt}\cdot\mathbf{g} + \mathbf{f}\cdot\dfrac{d\mathbf{g}}{dt}$

(g) $\dfrac{d}{dt}(\mathbf{f}\times\mathbf{g}) = \dfrac{d\mathbf{f}}{dt}\times\mathbf{g} + \mathbf{f}\times\dfrac{d\mathbf{g}}{dt}$

Proof: The proofs of parts (a)-(e) follow easily by differentiating the component functions and using the rules for derivatives from single-variable calculus. We will prove part (f), and leave the proof of part (g) as an exercise for the reader.

(f) Write $\mathbf{f}(t) = (f_1(t), f_2(t), f_3(t))$ and $\mathbf{g}(t) = (g_1(t), g_2(t), g_3(t))$, where the component functions $f_1(t)$, $f_2(t)$, $f_3(t)$, $g_1(t)$, $g_2(t)$, $g_3(t)$ are all differentiable real-valued functions. Then

$$\begin{aligned}
\frac{d}{dt}(\mathbf{f}(t)\cdot\mathbf{g}(t)) &= \frac{d}{dt}(f_1(t)g_1(t) + f_2(t)g_2(t) + f_3(t)g_3(t)) \\
&= \frac{d}{dt}(f_1(t)g_1(t)) + \frac{d}{dt}(f_2(t)g_2(t)) + \frac{d}{dt}(f_3(t)g_3(t)) \\
&= \frac{df_1}{dt}(t)g_1(t) + f_1(t)\frac{dg_1}{dt}(t) + \frac{df_2}{dt}(t)g_2(t) + f_2(t)\frac{dg_2}{dt}(t) + \frac{df_3}{dt}(t)g_3(t) + f_3(t)\frac{dg_3}{dt}(t) \\
&= \left(\frac{df_1}{dt}(t), \frac{df_2}{dt}(t), \frac{df_3}{dt}(t)\right)\cdot(g_1(t), g_2(t), g_3(t)) \\
&\quad + (f_1(t), f_2(t), f_3(t))\cdot\left(\frac{dg_1}{dt}(t), \frac{dg_2}{dt}(t), \frac{dg_3}{dt}(t)\right) \\
&= \frac{d\mathbf{f}}{dt}(t)\cdot\mathbf{g}(t) + \mathbf{f}(t)\cdot\frac{d\mathbf{g}}{dt}(t) \ \text{ for all } t.
\end{aligned}$$

QED

Example 1.37. Suppose $\mathbf{f}(t)$ is differentiable. Find the derivative of $\|\mathbf{f}(t)\|$.

Solution: Since $\|\mathbf{f}(t)\|$ is a real-valued function of t, then by the Chain Rule for real-valued functions, we know that $\frac{d}{dt}\|\mathbf{f}(t)\|^2 = 2\|\mathbf{f}(t)\|\frac{d}{dt}\|\mathbf{f}(t)\|$.

But $\|\mathbf{f}(t)\|^2 = \mathbf{f}(t)\cdot\mathbf{f}(t)$, so $\frac{d}{dt}\|\mathbf{f}(t)\|^2 = \frac{d}{dt}(\mathbf{f}(t)\cdot\mathbf{f}(t))$. Hence, we have

$$\begin{aligned} 2\|\mathbf{f}(t)\|\frac{d}{dt}\|\mathbf{f}(t)\| &= \frac{d}{dt}(\mathbf{f}(t)\cdot\mathbf{f}(t)) = \mathbf{f}'(t)\cdot\mathbf{f}(t) + \mathbf{f}(t)\cdot\mathbf{f}'(t) \text{ by Theorem 1.20(f), so} \\ &= 2\mathbf{f}'(t)\cdot\mathbf{f}(t)\text{ , so if } \|\mathbf{f}(t)\| \neq 0 \text{ then} \\ \frac{d}{dt}\|\mathbf{f}(t)\| &= \frac{\mathbf{f}'(t)\cdot\mathbf{f}(t)}{\|\mathbf{f}(t)\|}. \end{aligned}$$

We know that $\|\mathbf{f}(t)\|$ is constant if and only if $\frac{d}{dt}\|\mathbf{f}(t)\| = 0$ for all t. Also, $\mathbf{f}(t) \perp \mathbf{f}'(t)$ if and only if $\mathbf{f}'(t)\cdot\mathbf{f}(t) = 0$. Thus, the above example shows this important fact:

If $\|\mathbf{f}(t)\| \neq 0$, then $\|\mathbf{f}(t)\|$ is constant if and only if $\mathbf{f}(t) \perp \mathbf{f}'(t)$ for all t.

This means that if a curve lies completely on a sphere (or circle) centered at the origin, then the tangent vector $\mathbf{f}'(t)$ is always perpendicular to the *position vector* $\mathbf{f}(t)$.

Example 1.38. The *spherical spiral* $\mathbf{f}(t) = \left(\frac{\cos t}{\sqrt{1+a^2t^2}}, \frac{\sin t}{\sqrt{1+a^2t^2}}, \frac{-at}{\sqrt{1+a^2t^2}}\right)$, for $a \neq 0$.

Figure 1.8.3 shows the graph of the curve when $a = 0.2$. In the exercises, the reader will be asked to show that this curve lies on the sphere $x^2 + y^2 + z^2 = 1$ and to verify directly that $\mathbf{f}'(t)\cdot\mathbf{f}(t) = 0$ for all t.

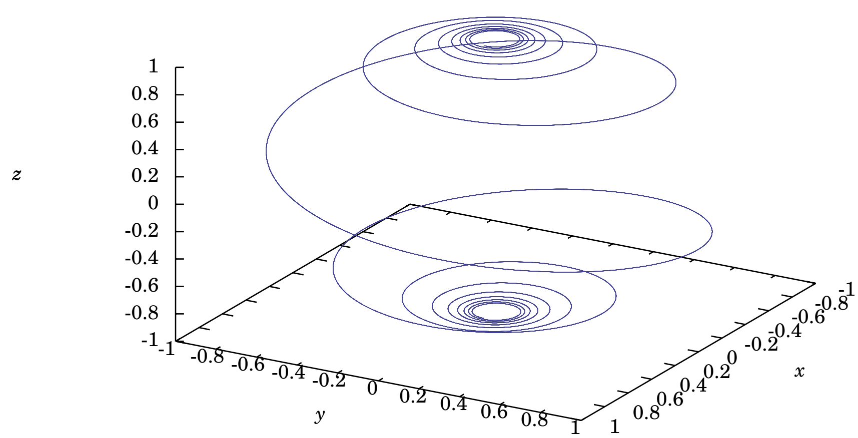


Figure 1.8.3 Spherical spiral with $a = 0.2$

Just as in single-variable calculus, higher-order derivatives of vector-valued functions are obtained by repeatedly differentiating the (first) derivative of the function:

$$\mathbf{f}''(t) = \frac{d}{dt}\mathbf{f}'(t)\,, \qquad \mathbf{f}'''(t) = \frac{d}{dt}\mathbf{f}''(t)\,, \quad \ldots \quad , \quad \frac{d^n\mathbf{f}}{dt^n} = \frac{d}{dt}\left(\frac{d^{n-1}\mathbf{f}}{dt^{n-1}}\right) \quad (\text{for } n = 2,3,4,\ldots)$$

We can use vector-valued functions to represent physical quantities, such as velocity, acceleration, force, momentum, etc. For example, let the real variable t represent time elapsed from some initial time ($t = 0$), and suppose that an object of constant mass m is subjected to some force so that it moves in space, with its position (x, y, z) at time t a function of t. That is, $x = x(t)$, $y = y(t)$, $z = z(t)$ for some real-valued functions $x(t)$, $y(t)$, $z(t)$. Call $\mathbf{r}(t) = (x(t), y(t), z(t))$ the **position vector** of the object. We can define various physical quantities associated with the object as follows:[14]

$$\begin{aligned}
\textit{position}\colon\ \mathbf{r}(t) &= (x(t), y(t), z(t))\\
\textit{velocity}\colon\ \mathbf{v}(t) &= \dot{\mathbf{r}}(t) = \mathbf{r}'(t) = \frac{d\mathbf{r}}{dt}\\
&= (x'(t), y'(t), z'(t))\\
\textit{acceleration}\colon\ \mathbf{a}(t) &= \dot{\mathbf{v}}(t) = \mathbf{v}'(t) = \frac{d\mathbf{v}}{dt}\\
&= \ddot{\mathbf{r}}(t) = \mathbf{r}''(t) = \frac{d^2\mathbf{r}}{dt^2}\\
&= (x''(t), y''(t), z''(t))\\
\textit{momentum}\colon\ \mathbf{p}(t) &= m\mathbf{v}(t)\\
\textit{force}\colon\ \mathbf{F}(t) &= \dot{\mathbf{p}}(t) = \mathbf{p}'(t) = \frac{d\mathbf{p}}{dt} \quad \text{(Newton's Second Law of Motion)}
\end{aligned}$$

The magnitude $\|\mathbf{v}(t)\|$ of the velocity vector is called the *speed* of the object. Note that since the mass m is a constant, the force equation becomes the familiar $\mathbf{F}(t) = m\mathbf{a}(t)$.

Example 1.39. Let $\mathbf{r}(t) = (5\cos t, 3\sin t, 4\sin t)$ be the position vector of an object at time $t \geq 0$. Find its (a) velocity and (b) acceleration vectors.

Solution: (a) $\mathbf{v}(t) = \dot{\mathbf{r}}(t) = (-5\sin t, 3\cos t, 4\cos t)$

(b) $\mathbf{a}(t) = \dot{\mathbf{v}}(t) = (-5\cos t, -3\sin t, -4\sin t)$

Note that $\|\mathbf{r}(t)\| = \sqrt{25\cos^2 t + 25\sin^2 t} = 5$ for all t, so by Example 1.37 we know that $\mathbf{r}(t) \cdot \dot{\mathbf{r}}(t) = 0$ for all t (which we can verify from part (a)). In fact, $\|\mathbf{v}(t)\| = 5$ for all t also. And not only does $\mathbf{r}(t)$ lie on the sphere of radius 5 centered at the origin, but perhaps not so obvious is that it lies completely within a *circle* of radius 5 centered at the origin. Also, note that $\mathbf{a}(t) = -\mathbf{r}(t)$. It turns out (see Exercise 16) that whenever an object moves in a circle with constant speed, the acceleration vector will point in the opposite direction of the position vector (i.e. towards the center of the circle).

---

[14] We will often use the older dot notation for derivatives when physics is involved.

Recall from Section 1.5 that if $\mathbf{r}_1$, $\mathbf{r}_2$ are position vectors to distinct points then $\mathbf{r}_1 + t(\mathbf{r}_2 - \mathbf{r}_1)$ represents a line through those two points as t varies over all real numbers. That vector sum can be written as $(1-t)\mathbf{r}_1 + t\mathbf{r}_2$. So the function $\mathbf{l}(t) = (1-t)\mathbf{r}_1 + t\mathbf{r}_2$ is a line through the terminal points of $\mathbf{r}_1$ and $\mathbf{r}_2$, and when t is restricted to the interval $[0,1]$ it is the line segment between the points, with $\mathbf{l}(0) = \mathbf{r}_1$ and $\mathbf{l}(1) = \mathbf{r}_2$.

In general, a function of the form $\mathbf{f}(t) = (a_1 t + b_1, a_2 t + b_2, a_3 t + b_3)$ represents a line in $\mathbb{R}^3$. A function of the form $\mathbf{f}(t) = (a_1 t^2 + b_1 t + c_1, a_2 t^2 + b_2 t + c_2, a_3 t^2 + b_3 t + c_3)$ represents a (possibly degenerate) parabola in $\mathbb{R}^3$.

Example 1.40. *Bézier curves* are used in Computer Aided Design (CAD) to approximate the shape of a polygonal path in space (called the *Bézier polygon* or *control polygon*). For instance, given three points (or position vectors) $\mathbf{b}_0$, $\mathbf{b}_1$, $\mathbf{b}_2$ in $\mathbb{R}^3$, define

$$\begin{aligned}
\mathbf{b}_0^1(t) &= (1-t)\mathbf{b}_0 + t\mathbf{b}_1 \\
\mathbf{b}_1^1(t) &= (1-t)\mathbf{b}_1 + t\mathbf{b}_2 \\
\mathbf{b}_0^2(t) &= (1-t)\mathbf{b}_0^1(t) + t\mathbf{b}_1^1(t) \\
&= (1-t)^2\mathbf{b}_0 + 2t(1-t)\mathbf{b}_1 + t^2\mathbf{b}_2
\end{aligned}$$

for all real t. For t in the interval $[0,1]$, we see that $\mathbf{b}_0^1(t)$ is the line segment between $\mathbf{b}_0$ and $\mathbf{b}_1$, and $\mathbf{b}_1^1(t)$ is the line segment between $\mathbf{b}_1$ and $\mathbf{b}_2$. The function $\mathbf{b}_0^2(t)$ is the Bézier curve for the points $\mathbf{b}_0$, $\mathbf{b}_1$, $\mathbf{b}_2$. Note from the last formula that the curve is a parabola that goes through $\mathbf{b}_0$ (when $t = 0$) and $\mathbf{b}_2$ (when $t = 1$).

As an example, let $\mathbf{b}_0 = (0,0,0)$, $\mathbf{b}_1 = (1,2,3)$, and $\mathbf{b}_2 = (4,5,2)$. Then the explicit formula for the Bézier curve is $\mathbf{b}_0^2(t) = (2t + 2t^2, 4t + t^2, 6t - 4t^2)$, as shown in Figure 1.8.4, where the line segments are $\mathbf{b}_0^1(t)$ and $\mathbf{b}_1^1(t)$, and the curve is $\mathbf{b}_0^2(t)$.

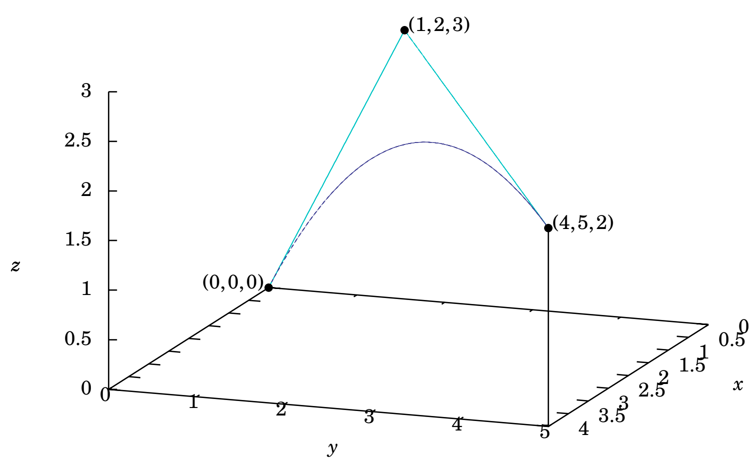


Figure 1.8.4 Bézier curve approximation for three points

In general, the polygonal path determined by $n \geq 3$ noncollinear points in $\mathbb{R}^3$ can be used to define the Bézier curve recursively by a process called *repeated linear interpolation*. This curve will be a vector-valued function whose components are polynomials of degree $n-1$, and its formula is given by *de Casteljau's algorithm*.[15] In the exercises, the reader will be given the algorithm for the case of $n = 4$ points and asked to write the explicit formula for the Bézier curve for the four points shown in Figure 1.8.5.

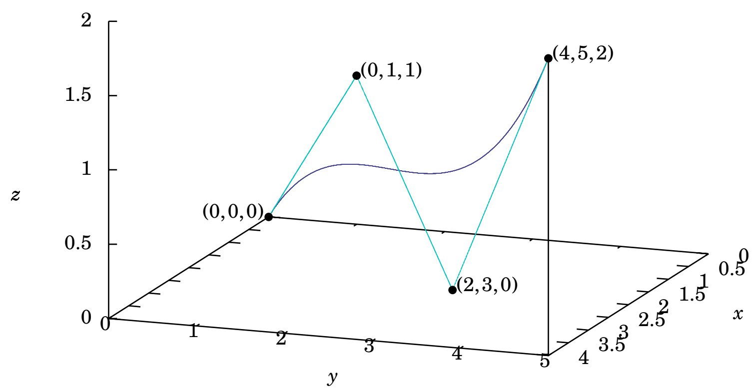


Figure 1.8.5 Bézier curve approximation for four points

Exercises

A

For Exercises 1-4, calculate $\mathbf{f}'(t)$ and find the tangent line at $\mathbf{f}(0)$.

1. $\mathbf{f}(t) = (t+1, t^2+1, t^3+1)$

2. $\mathbf{f}(t) = (e^t+1, e^{2t}+1, e^{t^2}+1)$

3. $\mathbf{f}(t) = (\cos 2t, \sin 2t, t)$

4. $\mathbf{f}(t) = (\sin 2t, 2\sin^2 t, 2\cos t)$

For Exercises 5-6, find the velocity $\mathbf{v}(t)$ and acceleration $\mathbf{a}(t)$ of an object with the given position vector $\mathbf{r}(t)$.

5. $\mathbf{r}(t) = (t, t - \sin t, 1 - \cos t)$

6. $\mathbf{r}(t) = (3\cos t, 2\sin t, 1)$

B

7. Let $\mathbf{f}(t) = \left(\dfrac{\cos t}{\sqrt{1+a^2t^2}}, \dfrac{\sin t}{\sqrt{1+a^2t^2}}, \dfrac{-at}{\sqrt{1+a^2t^2}}\right)$, with $a \neq 0$.

(a) Show that $\|\mathbf{f}(t)\| = 1$ for all t.

(b) Show directly that $\mathbf{f}'(t) \cdot \mathbf{f}(t) = 0$ for all t.

8. If $\mathbf{f}'(t) = \mathbf{0}$ for all t in some interval (a, b), show that $\mathbf{f}(t)$ is a constant vector in (a, b).

---

[15] See pp. 27-30 in FARIN.

9. For a constant vector $\mathbf{c} \neq \mathbf{0}$, the function $\mathbf{f}(t) = t\mathbf{c}$ represents a line parallel to $\mathbf{c}$.

(a) What kind of curve does $\mathbf{g}(t) = t^3\mathbf{c}$ represent? Explain.

(b) What kind of curve does $\mathbf{h}(t) = e^t\mathbf{c}$ represent? Explain.

(c) Compare $\mathbf{f}'(0)$ and $\mathbf{g}'(0)$. Given your answer to part (a), how do you explain the difference in the two derivatives?

10. Show that $\dfrac{d}{dt}\left(\mathbf{f} \times \dfrac{d\mathbf{f}}{dt}\right) = \mathbf{f} \times \dfrac{d^2\mathbf{f}}{dt^2}$.

11. Let a particle of (constant) mass m have position vector $\mathbf{r}(t)$, velocity $\mathbf{v}(t)$, acceleration $\mathbf{a}(t)$ and momentum $\mathbf{p}(t)$ at time t. The *angular momentum* $\mathbf{L}(t)$ of the particle with respect to the origin at time t is defined as $\mathbf{L}(t) = \mathbf{r}(t) \times \mathbf{p}(t)$. If $\mathbf{F}(t)$ is the force acting on the particle at time t, then define the *torque* $\mathbf{N}(t)$ acting on the particle with respect to the origin as $\mathbf{N}(t) = \mathbf{r}(t) \times \mathbf{F}(t)$. Show that $\mathbf{L}'(t) = \mathbf{N}(t)$.

12. Show that $\dfrac{d}{dt}(\mathbf{f} \cdot (\mathbf{g} \times \mathbf{h})) = \dfrac{d\mathbf{f}}{dt} \cdot (\mathbf{g} \times \mathbf{h}) + \mathbf{f} \cdot \left(\dfrac{d\mathbf{g}}{dt} \times \mathbf{h}\right) + \mathbf{f} \cdot \left(\mathbf{g} \times \dfrac{d\mathbf{h}}{dt}\right)$.

13. The Mean Value Theorem does not hold for vector-valued functions: Show that for $\mathbf{f}(t) = (\cos t, \sin t, t)$, there is no t in the interval $(0, 2\pi)$ such that

$$\mathbf{f}'(t) = \frac{\mathbf{f}(2\pi) - \mathbf{f}(0)}{2\pi - 0}.$$

C

14. The Bézier curve $\mathbf{b}_0^3(t)$ for four noncollinear points $\mathbf{b}_0$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ in $\mathbb{R}^3$ is defined by the following algorithm (going from the left column to the right):

$$\begin{array}{lll}
\mathbf{b}_0^1(t) = (1-t)\mathbf{b}_0 + t\mathbf{b}_1 & \mathbf{b}_0^2(t) = (1-t)\mathbf{b}_0^1(t) + t\mathbf{b}_1^1(t) & \mathbf{b}_0^3(t) = (1-t)\mathbf{b}_0^2(t) + t\mathbf{b}_1^2(t) \\
\mathbf{b}_1^1(t) = (1-t)\mathbf{b}_1 + t\mathbf{b}_2 & \mathbf{b}_1^2(t) = (1-t)\mathbf{b}_1^1(t) + t\mathbf{b}_2^1(t) & \\
\mathbf{b}_2^1(t) = (1-t)\mathbf{b}_2 + t\mathbf{b}_3 & &
\end{array}$$

(a) Show that $\mathbf{b}_0^3(t) = (1-t)^3\mathbf{b}_0 + 3t(1-t)^2\mathbf{b}_1 + 3t^2(1-t)\mathbf{b}_2 + t^3\mathbf{b}_3$.

(b) Write the explicit formula (as in Example 1.40) for the Bézier curve for the points $\mathbf{b}_0 = (0,0,0)$, $\mathbf{b}_1 = (0,1,1)$, $\mathbf{b}_2 = (2,3,0)$, $\mathbf{b}_3 = (4,5,2)$.

15. Let $\mathbf{r}(t)$ be the position vector for a particle moving in $\mathbb{R}^3$. Show that

$$\frac{d}{dt}(\mathbf{r} \times (\mathbf{v} \times \mathbf{r})) = \|\mathbf{r}\|^2\mathbf{a} + (\mathbf{r} \cdot \mathbf{v})\mathbf{v} - (\|\mathbf{v}\|^2 + \mathbf{r} \cdot \mathbf{a})\mathbf{r}.$$

16. Let $\mathbf{r}(t)$ be the position vector in $\mathbb{R}^3$ for a particle that moves with constant speed $c > 0$ in a circle of radius $a > 0$ in the xy-plane. Show that $\mathbf{a}(t)$ points in the opposite direction as $\mathbf{r}(t)$ for all t. (*Hint: Use Example 1.37 to show that* $\mathbf{r}(t) \perp \mathbf{v}(t)$ *and* $\mathbf{a}(t) \perp \mathbf{v}(t)$, *and hence* $\mathbf{a}(t) \parallel \mathbf{r}(t)$.)

17. Prove Theorem 1.20(g).

1.9 Arc Length

Let $\mathbf{r}(t) = (x(t), y(t), z(t))$ be the position vector of an object moving in $\mathbb{R}^3$. Since $\|\mathbf{v}(t)\|$ is the speed of the object at time t, it seems natural to define the distance s traveled by the object from time $t = a$ to $t = b$ as the definite integral

$$s = \int_a^b \|\mathbf{v}(t)\|\,dt = \int_a^b \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\,dt\,, \tag{1.40}$$

which is analogous to the case from single-variable calculus for parametric functions in $\mathbb{R}^2$. This is indeed how we will define the distance traveled and, in general, the arc length of a curve in $\mathbb{R}^3$.

Definition 1.13. Let $\mathbf{f}(t) = (x(t), y(t), z(t))$ be a curve in $\mathbb{R}^3$ whose domain includes the interval $[a,b]$. Suppose that in the interval (a,b) the first derivative of each component function $x(t)$, $y(t)$ and $z(t)$ exists and is continuous, and that no section of the curve is repeated. Then the **arc length** L of the curve from $t = a$ to $t = b$ is

$$L = \int_a^b \|\mathbf{f}'(t)\|\,dt = \int_a^b \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\,dt \tag{1.41}$$

A real-valued function whose first derivative is continuous is called *continuously differentiable* (or a $\mathscr{C}^1$ function), and a function whose derivatives of all orders are continuous is called *smooth* (or a $\mathscr{C}^\infty$ function). All the functions we will consider will be smooth. A *smooth curve* $\mathbf{f}(t)$ is one whose derivative $\mathbf{f}'(t)$ is never the zero vector and whose component functions are all smooth.

Note that we did not *prove* that the formula in the above definition actually gives the length of a section of a curve. A rigorous proof requires dealing with some subtleties, normally glossed over in calculus texts, which are beyond the scope of this book.[16]

Example 1.41. Find the length L of the helix $\mathbf{f}(t) = (\cos t, \sin t, t)$ from $t = 0$ to $t = 2\pi$.

Solution: By formula (1.41), we have

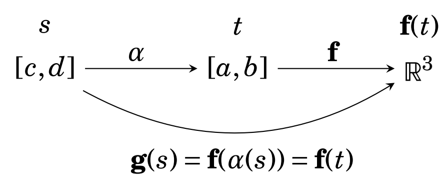
$$\begin{aligned} L &= \int_0^{2\pi} \sqrt{(-\sin t)^2 + (\cos t)^2 + 1^2}\,dt = \int_0^{2\pi} \sqrt{\sin^2 t + \cos^2 t + 1}\,dt = \int_0^{2\pi} \sqrt{2}\,dt \\ &= \sqrt{2}(2\pi - 0) = 2\sqrt{2}\pi \end{aligned}$$

Similar to the case in $\mathbb{R}^2$, if there are values of t in the interval $[a,b]$ where the derivative of a component function is not continuous then it is often possible to partition $[a,b]$ into subintervals where all the component functions are continuously differentiable (except at the endpoints, which can be ignored). The sum of the arc lengths over the subintervals will be the arc length over $[a,b]$.

[16] In particular, *Duhamel's principle* is needed. See the proof in Taylor and Mann, § 14.2 and § 18.2.

Notice that the curve traced out by the function $\mathbf{f}(t) = (\cos t, \sin t, t)$ from Example 1.41 is also traced out by the function $\mathbf{g}(t) = (\cos 2t, \sin 2t, 2t)$. For example, over the interval $[0,\pi]$, $\mathbf{g}(t)$ traces out the same section of the curve as $\mathbf{f}(t)$ does over the interval $[0,2\pi]$. Intuitively, this says that $\mathbf{g}(t)$ traces the curve twice as fast as $\mathbf{f}(t)$. This makes sense since, viewing the functions as position vectors and their derivatives as velocity vectors, the speeds of $\mathbf{f}(t)$ and $\mathbf{g}(t)$ are $\|\mathbf{f}'(t)\| = \sqrt{2}$ and $\|\mathbf{g}'(t)\| = 2\sqrt{2}$, respectively. We say that $\mathbf{g}(t)$ and $\mathbf{f}(t)$ are different *parametrizations* of the same curve.

Definition 1.14. Let C be a smooth curve in $\mathbb{R}^3$ represented by a function $\mathbf{f}(t)$ defined on an interval $[a,b]$, and let $\alpha : [c,d] \to [a,b]$ be a smooth one-to-one mapping of an interval $[c,d]$ onto $[a,b]$. Then the function $\mathbf{g} : [c,d] \to \mathbb{R}^3$ defined by $\mathbf{g}(s) = \mathbf{f}(\alpha(s))$ is a **parametrization** of C with **parameter** s. If α is strictly increasing on $[c,d]$ then we say that $\mathbf{g}(s)$ is *equivalent* to $\mathbf{f}(t)$.



Note that the differentiability of $\mathbf{g}(s)$ follows from a version of the Chain Rule for vector-valued functions (the proof is left as an exercise):

Theorem 1.21. Chain Rule: If $\mathbf{f}(t)$ is a differentiable vector-valued function of t, and $t = \alpha(s)$ is a differentiable scalar function of s, then $\mathbf{f}(s) = \mathbf{f}(\alpha(s))$ is a differentiable vector-valued function of s, and

$$\frac{d\mathbf{f}}{ds} = \frac{d\mathbf{f}}{dt}\frac{dt}{ds} \tag{1.42}$$

for any s where the composite function $\mathbf{f}(\alpha(s))$ is defined.

Example 1.42. The following are all equivalent parametrizations of the same curve:

$$\mathbf{f}(t) = (\cos t, \sin t, t) \quad \text{for } t \text{ in } [0,2\pi]$$
$$\mathbf{g}(s) = (\cos 2s, \sin 2s, 2s) \quad \text{for } s \text{ in } [0,\pi]$$
$$\mathbf{h}(s) = (\cos 2\pi s, \sin 2\pi s, 2\pi s) \quad \text{for } s \text{ in } [0,1]$$

To see that $\mathbf{g}(s)$ is equivalent to $\mathbf{f}(t)$, define $\alpha : [0,\pi] \to [0,2\pi]$ by $\alpha(s) = 2s$. Then α is smooth, one-to-one, maps $[0,\pi]$ onto $[0,2\pi]$, and is strictly increasing (since $\alpha'(s) = 2 > 0$ for all s). Likewise, defining $\alpha : [0,1] \to [0,2\pi]$ by $\alpha(s) = 2\pi s$ shows that $\mathbf{h}(s)$ is equivalent to $\mathbf{f}(t)$.

A curve can have many parametrizations, with different speeds, so which one is the best to use? In some situations the **arc length parametrization** can be useful. The idea behind this is to replace the parameter t, for any given smooth parametrization $\mathbf{f}(t)$ defined on $[a,b]$, by the parameter s given by

$$s = s(t) = \int_a^t \|\mathbf{f}'(u)\|\,du. \tag{1.43}$$

In terms of motion along a curve, s is the distance traveled along the curve after time t has elapsed. So the new parameter will be distance instead of time. There is a natural correspondence between s and t: from a starting point on the curve, the distance traveled along the curve (in one direction) is uniquely determined by the amount of time elapsed, and vice versa.

Since s is the arc length of the curve over the interval $[a,t]$ for each t in $[a,b]$, then it is a function of t. By the Fundamental Theorem of Calculus, its derivative is

$$s'(t) = \frac{ds}{dt} = \frac{d}{dt}\int_a^t \|\mathbf{f}'(u)\|\,du = \|\mathbf{f}'(t)\| \quad \text{for all } t \text{ in } [a,b].$$

Since $\mathbf{f}(t)$ is smooth, then $\|\mathbf{f}'(t)\| > 0$ for all t in $[a,b]$. Thus $s'(t) > 0$ and hence $s(t)$ is strictly increasing on the interval $[a,b]$. Recall that this means that s is a one-to-one mapping of the interval $[a,b]$ onto the interval $[s(a), s(b)]$. But we see that

$$s(a) = \int_a^a \|\mathbf{f}'(u)\|\,du = 0 \quad \text{and} \quad s(b) = \int_a^b \|\mathbf{f}'(u)\|\,du = L = \text{arc length from } t = a \text{ to } t = b$$

So the function $s : [a,b] \to [0,L]$ is a one-to-one, differentiable mapping onto the interval $[0,L]$. From single-variable calculus, we know that this means that there exists an inverse function $\alpha : [0,L] \to [a,b]$ that is differentiable and the inverse of $s : [a,b] \to [0,L]$. That is, for each t in $[a,b]$ there is a unique s in $[0,L]$ such that $s = s(t)$ and $t = \alpha(s)$. And we know that the derivative of α is

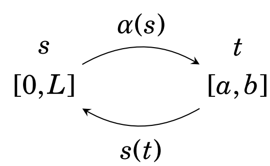


Figure 1.9.1 $t = \alpha(s)$

$$\alpha'(s) = \frac{1}{s'(\alpha(s))} = \frac{1}{\|\mathbf{f}'(\alpha(s))\|}$$

So define the arc length parametrization $\mathbf{f} : [0,L] \to \mathbb{R}^3$ by

$$\mathbf{f}(s) = \mathbf{f}(\alpha(s)) \quad \text{for all } s \text{ in } [0,L].$$

Then $\mathbf{f}(s)$ is smooth, by the Chain Rule. In fact, $\mathbf{f}(s)$ has *unit speed*:

$$\begin{aligned}
\mathbf{f}'(s) &= \mathbf{f}'(\alpha(s))\,\alpha'(s) \quad \text{by the Chain Rule, so} \\
&= \mathbf{f}'(\alpha(s))\,\frac{1}{\|\mathbf{f}'(\alpha(s))\|} \quad \text{, so} \\
\|\mathbf{f}'(s)\| &= 1 \quad \text{for all } s \text{ in } [0,L].
\end{aligned}$$

So the arc length parametrization traverses the curve at a “normal” rate.

In practice, parametrizing a curve $\mathbf{f}(t)$ by arc length requires you to evaluate the integral $s = \int_a^t \|\mathbf{f}'(u)\|\,du$ in some closed form (as a function of t) so that you could then solve for t in terms of s. If that can be done, you would then substitute the expression for t in terms of s (which we called $\alpha(s)$) into the formula for $\mathbf{f}(t)$ to get $\mathbf{f}(s)$.

Example 1.43. Parametrize the helix $\mathbf{f}(t) = (\cos t, \sin t, t)$, for t in $[0, 2\pi]$, by arc length.

Solution: By Example 1.41 and formula (1.43), we have

$$s = \int_0^t \|\mathbf{f}'(u)\|\,du = \int_0^t \sqrt{2}\,du = \sqrt{2}\,t \quad \text{for all } t \text{ in } [0, 2\pi].$$

So we can solve for t in terms of s: $t = \alpha(s) = \dfrac{s}{\sqrt{2}}$.

$\therefore\ \ \mathbf{f}(s) = \left(\cos \dfrac{s}{\sqrt{2}}, \sin \dfrac{s}{\sqrt{2}}, \dfrac{s}{\sqrt{2}}\right)$ for all s in $[0, 2\sqrt{2}\pi]$. Note that $\|\mathbf{f}'(s)\| = 1$.

Arc length plays an important role when discussing *curvature* and *moving frame fields*, in the field of mathematics known as *differential geometry*.[17] The methods involve using an arc length parametrization, which often leads to an integral that is either difficult or impossible to evaluate in a simple closed form. The simple integral in Example 1.43 is the exception, not the norm. In general, arc length parametrizations are more useful for theoretical purposes than for practical computations.[18] Curvature and moving frame fields can be defined without using arc length, which makes their computation much easier, and these definitions can be shown to be equivalent to those using arc length. We will leave this to the exercises.

The arc length for curves given in other coordinate systems can also be calculated:

Theorem 1.22. Suppose that $r = r(t)$, $\theta = \theta(t)$ and $z = z(t)$ are the cylindrical coordinates of a curve $\mathbf{f}(t)$, for t in $[a, b]$. Then the arc length L of the curve over $[a, b]$ is

$$L = \int_a^b \sqrt{r'(t)^2 + r(t)^2\theta'(t)^2 + z'(t)^2}\,dt \tag{1.44}$$

Proof: The Cartesian coordinates $(x(t), y(t), z(t))$ of a point on the curve are given by

$$x(t) = r(t)\cos\theta(t), \quad y(t) = r(t)\sin\theta(t), \quad z(t) = z(t)$$

so differentiating the above expressions for $x(t)$ and $y(t)$ with respect to t gives

$$x'(t) = r'(t)\cos\theta(t) - r(t)\theta'(t)\sin\theta(t), \quad y'(t) = r'(t)\sin\theta(t) + r(t)\theta'(t)\cos\theta(t)$$

[17] See O'NEILL for an introduction to elementary differential geometry.

[18] For example, the usual parametrizations of Bézier curves, which we discussed in Section 1.8, are polynomial functions in $\mathbb{R}^3$. This makes their computation relatively simple, which, in CAD, is desirable. But their arc length parametrizations are not only *not* polynomials, they are in fact usually impossible to calculate at all.

and so

$$\begin{aligned}
x'(t)^2 + y'(t)^2 &= (r'(t)\cos\theta(t) - r(t)\theta'(t)\sin\theta(t))^2 + (r'(t)\sin\theta(t) + r(t)\theta'(t)\cos\theta(t))^2\\
&= r'(t)^2(\cos^2\theta + \sin^2\theta) + r(t)^2\theta'(t)^2(\cos^2\theta + \sin^2\theta)\\
&\quad - 2r'(t)r(t)\theta'(t)\cos\theta\sin\theta + 2r'(t)r(t)\theta'(t)\cos\theta\sin\theta\\
&= r'(t)^2 + r(t)^2\theta'(t)^2 \quad \text{, and so}\\
L &= \int_a^b \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\,dt\\
&= \int_a^b \sqrt{r'(t)^2 + r(t)^2\theta'(t)^2 + z'(t)^2}\,dt
\end{aligned}$$

QED

Example 1.44. Find the arc length L of the curve whose cylindrical coordinates are $r = e^t$, $\theta = t$ and $z = e^t$, for t over the interval $[0,1]$.

Solution: Since $r'(t) = e^t$, $\theta'(t) = 1$ and $z'(t) = e^t$, then

$$\begin{aligned}
L &= \int_0^1 \sqrt{r'(t)^2 + r(t)^2\theta'(t)^2 + z'(t)^2}\,dt\\
&= \int_0^1 \sqrt{e^{2t} + e^{2t}(1) + e^{2t}}\,dt\\
&= \int_0^1 e^t\sqrt{3}\,dt = \sqrt{3}(e-1)
\end{aligned}$$

Exercises

A

For Exercises 1-3, calculate the arc length of $\mathbf{f}(t)$ over the given interval.

1. $\mathbf{f}(t) = (3\cos 2t, 3\sin 2t, 3t)$ on $[0, \pi/2]$

2. $\mathbf{f}(t) = ((t^2+1)\cos t, (t^2+1)\sin t, 2\sqrt{2}t)$ on $[0,1]$

3. $\mathbf{f}(t) = (2\cos 3t, 2\sin 3t, 2t^{3/2})$ on $[0,1]$

4. Parametrize the curve from Exercise 1 by arc length.

5. Parametrize the curve from Exercise 3 by arc length.

B

6. Let $\mathbf{f}(t)$ be a differentiable curve such that $\mathbf{f}(t) \neq \mathbf{0}$ for all t. Show that

$$\frac{d}{dt}\left(\frac{\mathbf{f}(t)}{\|\mathbf{f}(t)\|}\right) = \frac{\mathbf{f}(t) \times (\mathbf{f}'(t) \times \mathbf{f}(t))}{\|\mathbf{f}(t)\|^3}.$$

Exercises 7-9 develop the moving frame field **T**, **N**, **B** at a point on a curve.

7. Let $\mathbf{f}(t)$ be a smooth curve such that $\mathbf{f}'(t) \neq \mathbf{0}$ for all t. Then we can define the *unit tangent vector* **T** by

$$\mathbf{T}(t) = \frac{\mathbf{f}'(t)}{\|\mathbf{f}'(t)\|}.$$

Show that

$$\mathbf{T}'(t) = \frac{\mathbf{f}'(t) \times (\mathbf{f}''(t) \times \mathbf{f}'(t))}{\|\mathbf{f}'(t)\|^3}.$$

8. Continuing Exercise 7, assume that $\mathbf{f}'(t)$ and $\mathbf{f}''(t)$ are not parallel. Then $\mathbf{T}'(t) \neq \mathbf{0}$ so we can define the *unit principal normal vector* **N** by

$$\mathbf{N}(t) = \frac{\mathbf{T}'(t)}{\|\mathbf{T}'(t)\|}.$$

Show that

$$\mathbf{N}(t) = \frac{\mathbf{f}'(t) \times (\mathbf{f}''(t) \times \mathbf{f}'(t))}{\|\mathbf{f}'(t)\| \, \|\mathbf{f}''(t) \times \mathbf{f}'(t)\|}.$$

9. Continuing Exercise 8, the *unit binormal vector* B is defined by

$$\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t).$$

Show that

$$\mathbf{B}(t) = \frac{\mathbf{f}'(t) \times \mathbf{f}''(t)}{\|\mathbf{f}'(t) \times \mathbf{f}''(t)\|}.$$

Note: The vectors $\mathbf{T}(t)$, $\mathbf{N}(t)$ and $\mathbf{B}(t)$ form a right-handed system of mutually perpendicular unit vectors (called *orthonormal vectors*) at each point on the curve $\mathbf{f}(t)$.

10. Continuing Exercise 9, the *curvature* κ is defined by

$$\kappa(t) = \frac{\|\mathbf{T}'(t)\|}{\|\mathbf{f}'(t)\|} = \frac{\|\mathbf{f}'(t) \times (\mathbf{f}''(t) \times \mathbf{f}'(t))\|}{\|\mathbf{f}'(t)\|^4}.$$

Show that

$$\kappa(t) = \frac{\|\mathbf{f}'(t) \times \mathbf{f}''(t)\|}{\|\mathbf{f}'(t)\|^3} \quad \text{and that} \quad \mathbf{T}'(t) = \|\mathbf{f}'(t)\| \, \kappa(t) \mathbf{N}(t).$$

Note: $\kappa(t)$ gives a sense of how "curved" the curve $\mathbf{f}(t)$ is at each point.

11. Find **T**, **N**, **B** and κ at each point of the helix $\mathbf{f}(t) = (\cos t, \sin t, t)$.

12. Show that the arc length L of a curve whose spherical coordinates are $\rho = \rho(t)$, $\theta = \theta(t)$ and $\phi = \phi(t)$ for t in an interval $[a, b]$ is

$$L = \int_a^b \sqrt{\rho'(t)^2 + (\rho(t)^2 \sin^2 \phi(t))\theta'(t)^2 + \rho(t)^2 \phi'(t)^2} \, dt.$$

2 Functions of Several Variables

2.1 Functions of Two or Three Variables

In Section 1.8 we discussed vector-valued functions of a single real variable. We will now examine real-valued functions of a point (or vector) in $\mathbb{R}^2$ or $\mathbb{R}^3$. For the most part these functions will be defined on sets of points in $\mathbb{R}^2$, but there will be times when we will use points in $\mathbb{R}^3$, and there will also be times when it will be convenient to think of the points as vectors (or terminal points of vectors).

A **real-valued function** f defined on a subset D of $\mathbb{R}^2$ is a rule that assigns to each point (x, y) in D a real number $f(x, y)$. The largest possible set D in $\mathbb{R}^2$ on which f is defined is called the **domain** of f, and the **range** of f is the set of all real numbers $f(x, y)$ as (x, y) varies over the domain D. A similar definition holds for functions $f(x, y, z)$ defined on points (x, y, z) in $\mathbb{R}^3$.

Example 2.1. The domain of the function

$$f(x, y) = xy$$

is all of $\mathbb{R}^2$, and the range of f is all of $\mathbb{R}$.

Example 2.2. The domain of the function

$$f(x, y) = \frac{1}{x - y}$$

is all of $\mathbb{R}^2$ except the points (x, y) for which $x = y$. That is, the domain is the set $D = \{(x, y) : x \neq y\}$. The range of f is all real numbers except 0.

Example 2.3. The domain of the function

$$f(x, y) = \sqrt{1 - x^2 - y^2}$$

is the set $D = \{(x, y) : x^2 + y^2 \le 1\}$, since the quantity inside the square root is nonnegative if and only if $1 - (x^2 + y^2) \ge 0$. We see that D consists of all points on and inside the unit circle in $\mathbb{R}^2$ (D is sometimes called the *closed unit disk*). The range of f is the interval $[0, 1]$ in $\mathbb{R}$.

Example 2.4. The domain of the function

$$f(x,y,z) = e^{x+y-z}$$

is all of $\mathbb{R}^3$, and the range of f is all positive real numbers.

A function $f(x,y)$ defined in $\mathbb{R}^2$ is often written as $z = f(x,y)$, as was mentioned in Section 1.1, so that the **graph** of $f(x,y)$ is the set $\{(x,y,z) : z = f(x,y)\}$ in $\mathbb{R}^3$. So we see that this graph is a surface in $\mathbb{R}^3$, since it satisfies an equation of the form $F(x,y,z) = 0$ (namely, $F(x,y,z) = f(x,y) - z$). The traces of this surface in the planes $z = c$, where c varies over $\mathbb{R}$, are called the **level curves** of the function. Equivalently, the level curves are the solution sets of the equations $f(x,y) = c$, for c in $\mathbb{R}$. Level curves are often projected onto the xy-plane to give an idea of the various "elevation" levels of the surface (as is done in topography).

Example 2.5. The graph of the function

$$f(x,y) = \frac{\sin\sqrt{x^2+y^2}}{\sqrt{x^2+y^2}}$$

is shown below. Note that the level curves (shown both on the surface and projected onto the xy-plane) are groups of concentric circles.

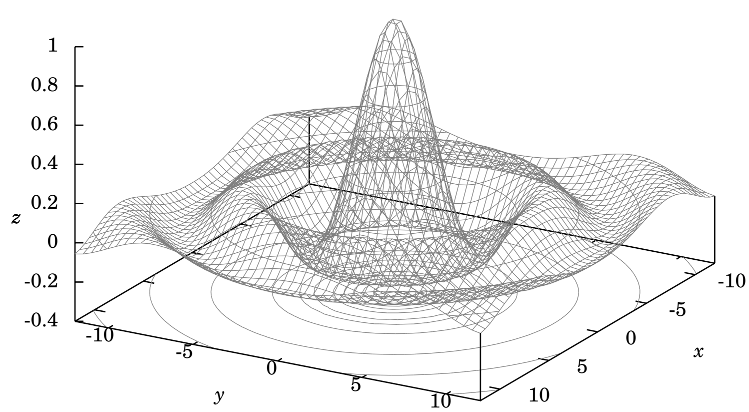


Figure 2.1.1 The function $f(x,y) = \frac{\sin\sqrt{x^2+y^2}}{\sqrt{x^2+y^2}}$

You may be wondering what happens to the function in Example 2.5 at the point $(x,y) = (0,0)$, since both the numerator and denominator are 0 at that point. The function is not defined at $(0,0)$, but the *limit* of the function exists (and equals 1) as (x,y) *approaches* $(0,0)$. We will now state explicitly what is meant by the limit of a function of two variables.

Definition 2.1. Let (a,b) be a point in $\mathbb{R}^2$, and let $f(x,y)$ be a real-valued function defined on some set containing (a,b) (but not necessarily defined at (a,b) itself). Then we say that the **limit** of $f(x,y)$ equals L as (x,y) approaches (a,b), written as

$$\lim_{(x,y)\to(a,b)} f(x,y) = L\,, \tag{2.1}$$

if given any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x,y) - L| < \epsilon \quad \text{whenever} \quad 0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta.$$

A similar definition can be made for functions of three variables. The idea behind the above definition is that the values of $f(x,y)$ can get arbitrarily close to L (i.e. within ϵ of L) if we pick (x,y) sufficiently close to (a,b) (i.e. inside a circle centered at (a,b) with some sufficiently small radius δ).

If you recall the "epsilon-delta" proofs of limits of real-valued functions of a single variable, you may remember how awkward they can be, and how they can usually only be done easily for simple functions. In general, the multivariable cases are at least equally awkward to go through, so we will not bother with such proofs. Instead, we will simply state that when the function $f(x,y)$ is given by a single formula and is defined at the point (a,b) (e.g. is not some indeterminate form like 0/0) then you can just substitute $(x,y) = (a,b)$ into the formula for $f(x,y)$ to find the limit.

Example 2.6.

$$\lim_{(x,y)\to(1,2)} \frac{xy}{x^2+y^2} = \frac{(1)(2)}{1^2+2^2} = \frac{2}{5}$$

since $f(x,y) = \frac{xy}{x^2+y^2}$ is properly defined at the point $(1,2)$.

The major difference between limits in one variable and limits in two or more variables has to do with how a point is approached. In the single-variable case, the statement "$x \to a$" means that x gets closer to the value a from two possible directions along the real number line (see Figure 2.1.2(a)). In two dimensions, however, (x,y) can approach a point (a,b) along an infinite number of paths (see Figure 2.1.2(b)).

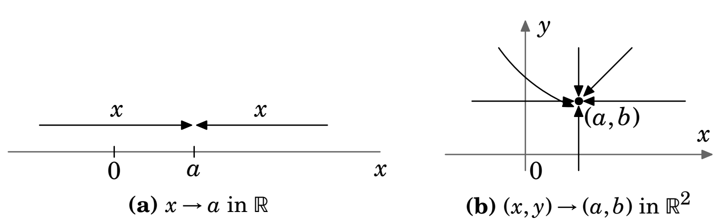


Figure 2.1.2 "Approaching" a point in different dimensions

Example 2.7.

$$\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2} \quad \text{does not exist}$$

Note that we can not simply substitute $(x, y) = (0,0)$ into the function, since doing so gives an indeterminate form 0/0. To show that the limit does not exist, we will show that the function approaches different values as (x, y) approaches $(0,0)$ along different paths in $\mathbb{R}^2$. To see this, suppose that $(x, y) \to (0,0)$ along the positive x-axis, so that $y = 0$ along that path. Then

$$f(x,y) = \frac{xy}{x^2+y^2} = \frac{x0}{x^2+0^2} = 0$$

along that path (since $x > 0$ in the denominator). But if $(x, y) \to (0,0)$ along the straight line $y = x$ through the origin, for $x > 0$, then we see that

$$f(x,y) = \frac{xy}{x^2+y^2} = \frac{x^2}{x^2+x^2} = \frac{1}{2}$$

which means that $f(x, y)$ approaches different values as $(x, y) \to (0,0)$ along different paths. Hence the limit does not exist.

Limits of real-valued multivariable functions obey the same algebraic rules as in the single-variable case, as shown in the following theorem, which we state without proof.

Theorem 2.1. Suppose that $\lim_{(x,y)\to(a,b)} f(x,y)$ and $\lim_{(x,y)\to(a,b)} g(x,y)$ both exist, and that k is some scalar. Then:

(a) $\displaystyle\lim_{(x,y)\to(a,b)} [f(x,y) \pm g(x,y)] = \left[\lim_{(x,y)\to(a,b)} f(x,y)\right] \pm \left[\lim_{(x,y)\to(a,b)} g(x,y)\right]$

(b) $\displaystyle\lim_{(x,y)\to(a,b)} k\,f(x,y) = k\left[\lim_{(x,y)\to(a,b)} f(x,y)\right]$

(c) $\displaystyle\lim_{(x,y)\to(a,b)} [f(x,y)g(x,y)] = \left[\lim_{(x,y)\to(a,b)} f(x,y)\right]\left[\lim_{(x,y)\to(a,b)} g(x,y)\right]$

(d) $\displaystyle\lim_{(x,y)\to(a,b)} \frac{f(x,y)}{g(x,y)} = \frac{\lim_{(x,y)\to(a,b)} f(x,y)}{\lim_{(x,y)\to(a,b)} g(x,y)} \quad \text{if} \lim_{(x,y)\to(a,b)} g(x,y) \neq 0$

(e) If $|f(x,y) - L| \le g(x,y)$ for all (x, y) and if $\lim_{(x,y)\to(a,b)} g(x,y) = 0$, then $\lim_{(x,y)\to(a,b)} f(x,y) = L$.

Note that in part (e), it suffices to have $|f(x,y) - L| \le g(x,y)$ for all (x, y) "sufficiently close" to (a, b) (but excluding (a, b) itself).

Example 2.8. Show that

$$\lim_{(x,y)\to(0,0)} \frac{y^4}{x^2+y^2} = 0.$$

Since substituting $(x,y) = (0,0)$ into the function gives the indeterminate form 0/0, we need an alternate method for evaluating this limit. We will use Theorem 2.1(e). First, notice that $y^4 = \left(\sqrt{y^2}\right)^4$ and so $0 \le y^4 \le \left(\sqrt{x^2+y^2}\right)^4$ for all (x,y). But $\left(\sqrt{x^2+y^2}\right)^4 = (x^2+y^2)^2$. Thus, for all $(x,y) \ne (0,0)$ we have

$$\left|\frac{y^4}{x^2+y^2}\right| \le \frac{(x^2+y^2)^2}{x^2+y^2} = x^2+y^2 \to 0 \quad \text{as} \quad (x,y)\to(0,0).$$

Therefore $\lim_{(x,y)\to(0,0)} \frac{y^4}{x^2+y^2} = 0$.

Continuity can be defined similarly as in the single-variable case.

Definition 2.2. A real-valued function $f(x,y)$ with domain D in $\mathbb{R}^2$ is **continuous** at the point (a,b) in D if $\lim_{(x,y)\to(a,b)} f(x,y) = f(a,b)$. We say that $f(x,y)$ is a **continuous function** if it is continuous at every point in its domain D.

Unless indicated otherwise, you can assume that all the functions we deal with are continuous. In fact, we can modify the function from Example 2.8 so that it is continuous on all of $\mathbb{R}^2$.

Example 2.9. Define a function $f(x,y)$ on all of $\mathbb{R}^2$ as follows:

$$f(x,y) = \begin{cases} 0 & \text{if } (x,y) = (0,0) \\ \dfrac{y^4}{x^2+y^2} & \text{if } (x,y) \ne (0,0) \end{cases}$$

Then $f(x,y)$ is well-defined for all (x,y) in $\mathbb{R}^2$ (i.e. there are no indeterminate forms for any (x,y)), and we see that

$$\lim_{(x,y)\to(a,b)} f(x,y) = \frac{b^4}{a^2+b^2} = f(a,b) \quad \text{for } (a,b) \ne (0,0).$$

So since

$$\lim_{(x,y)\to(0,0)} f(x,y) = 0 = f(0,0) \quad \text{by Example 2.8,}$$

then $f(x,y)$ is continuous on all of $\mathbb{R}^2$.

Exercises

A

For Exercises 1-6, state the domain and range of the given function.

1. $f(x,y) = x^2 + y^2 - 1$

2. $f(x,y) = \dfrac{1}{x^2 + y^2}$

3. $f(x,y) = \sqrt{x^2 + y^2 - 4}$

4. $f(x,y) = \dfrac{x^2 + 1}{y}$

5. $f(x,y,z) = \sin(xyz)$

6. $f(x,y,z) = \sqrt{(x-1)(yz-1)}$

For Exercises 7-18, evaluate the given limit.

7. $\displaystyle\lim_{(x,y)\to(0,0)} \cos(xy)$

8. $\displaystyle\lim_{(x,y)\to(0,0)} e^{xy}$

9. $\displaystyle\lim_{(x,y)\to(0,0)} \frac{x^2 - y^2}{x^2 + y^2}$

10. $\displaystyle\lim_{(x,y)\to(0,0)} \frac{xy^2}{x^2 + y^4}$

11. $\displaystyle\lim_{(x,y)\to(1,-1)} \frac{x^2 - 2xy + y^2}{x - y}$

12. $\displaystyle\lim_{(x,y)\to(0,0)} \frac{xy^2}{x^2 + y^2}$

13. $\displaystyle\lim_{(x,y)\to(1,1)} \frac{x^2 - y^2}{x - y}$

14. $\displaystyle\lim_{(x,y)\to(0,0)} \frac{x^2 - 2xy + y^2}{x - y}$

15. $\displaystyle\lim_{(x,y)\to(0,0)} \frac{y^4 \sin(xy)}{x^2 + y^2}$

16. $\displaystyle\lim_{(x,y)\to(0,0)} (x^2 + y^2)\cos\left(\frac{1}{xy}\right)$

17. $\displaystyle\lim_{(x,y)\to(0,0)} \frac{x}{y}$

18. $\displaystyle\lim_{(x,y)\to(0,0)} \cos\left(\frac{1}{xy}\right)$

B

19. Show that $f(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$, for $\sigma > 0$, is constant on the circle of radius $r > 0$ centered at the origin. This function is called a *Gaussian blur*, and is used as a filter in image processing software to produce a "blurred" effect.

20. Suppose that $f(x,y) \le f(y,x)$ for all (x,y) in $\mathbb{R}^2$. Show that $f(x,y) = f(y,x)$ for all (x,y) in $\mathbb{R}^2$.

21. Use the substitution $r = \sqrt{x^2 + y^2}$ to show that

$$\lim_{(x,y)\to(0,0)} \frac{\sin\sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2}} = 1\ .$$

(*Hint: You will need to use L'Hôpital's Rule for single-variable limits.*)

C

22. Prove Theorem 2.1(a) in the case of addition. (*Hint: Use Definition 2.1.*)

23. Prove Theorem 2.1(b).

2.2 Partial Derivatives

Now that we have an idea of what functions of several variables are, and what a limit of such a function is, we can start to develop an idea of a derivative of a function of two or more variables. We will start with the notion of a *partial derivative*.

Definition 2.3. Let $f(x,y)$ be a real-valued function with domain D in $\mathbb{R}^2$, and let (a,b) be a point in D. Then the **partial derivative of f at (a,b) with respect to x**, denoted by $\frac{\partial f}{\partial x}(a,b)$, is defined as

$$\frac{\partial f}{\partial x}(a,b) = \lim_{h\to 0} \frac{f(a+h,b)-f(a,b)}{h} \tag{2.2}$$

and the **partial derivative of f at (a,b) with respect to y**, denoted by $\frac{\partial f}{\partial y}(a,b)$, is defined as

$$\frac{\partial f}{\partial y}(a,b) = \lim_{h\to 0} \frac{f(a,b+h)-f(a,b)}{h} \; . \tag{2.3}$$

Note: The symbol ∂ is pronounced "del".[1]

Recall that the derivative of a function $f(x)$ can be interpreted as the rate of change of that function in the (positive) x direction. From the definitions above, we can see that the partial derivative of a function $f(x,y)$ with respect to x or y is the rate of change of $f(x,y)$ in the (positive) x or y direction, respectively. What this means is that the partial derivative of a function $f(x,y)$ with respect to x can be calculated by treating the y *variable* as a *constant*, and then simply differentiating $f(x,y)$ as if it were a function of x alone, using the usual rules from single-variable calculus. Likewise, the partial derivative of $f(x,y)$ with respect to y is obtained by treating the x variable as a constant and then differentiating $f(x,y)$ as if it were a function of y alone.

Example 2.10. Find $\frac{\partial f}{\partial x}(x,y)$ and $\frac{\partial f}{\partial y}(x,y)$ for the function $f(x,y) = x^2y + y^3$.

Solution: Treating y as a constant and differentiating $f(x,y)$ with respect to x gives

$$\frac{\partial f}{\partial x}(x,y) = 2xy$$

and treating x as a constant and differentiating $f(x,y)$ with respect to y gives

$$\frac{\partial f}{\partial y}(x,y) = x^2 + 3y^2 \; .$$

---

[1] It is not a Greek letter. The symbol was first used by the mathematicians A. Clairaut and L. Euler around 1740, to distinguish it from the letter d used for the "usual" derivative.

We will often simply write $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ instead of $\frac{\partial f}{\partial x}(x,y)$ and $\frac{\partial f}{\partial y}(x,y)$.

Example 2.11. Find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ for the function $f(x,y) = \frac{\sin(xy^2)}{x^2+1}$.

Solution: Treating y as a constant and differentiating $f(x,y)$ with respect to x gives

$$\frac{\partial f}{\partial x} = \frac{(x^2+1)(y^2\cos(xy^2)) - (2x)\sin(xy^2)}{(x^2+1)^2}$$

and treating x as a constant and differentiating $f(x,y)$ with respect to y gives

$$\frac{\partial f}{\partial y} = \frac{2xy\cos(xy^2)}{x^2+1}\ .$$

Since both $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are themselves functions of x and y, we can take *their* partial derivatives with respect to x and y. This yields the *higher-order partial derivatives*:

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) & \frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)\\
\frac{\partial^2 f}{\partial y\,\partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) & \frac{\partial^2 f}{\partial x\,\partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)\\
\frac{\partial^3 f}{\partial x^3} &= \frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial x^2}\right) & \frac{\partial^3 f}{\partial y^3} &= \frac{\partial}{\partial y}\left(\frac{\partial^2 f}{\partial y^2}\right)\\
\frac{\partial^3 f}{\partial y\,\partial x^2} &= \frac{\partial}{\partial y}\left(\frac{\partial^2 f}{\partial x^2}\right) & \frac{\partial^3 f}{\partial x\,\partial y^2} &= \frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial y^2}\right)\\
\frac{\partial^3 f}{\partial y^2\,\partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial^2 f}{\partial y\,\partial x}\right) & \frac{\partial^3 f}{\partial x^2\,\partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial x\,\partial y}\right)\\
\frac{\partial^3 f}{\partial x\,\partial y\,\partial x} &= \frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial y\,\partial x}\right) & \frac{\partial^3 f}{\partial y\,\partial x\,\partial y} &= \frac{\partial}{\partial y}\left(\frac{\partial^2 f}{\partial x\,\partial y}\right)\\
&\ \ \vdots
\end{aligned}$$

Example 2.12. Find the partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial y^2}$, $\frac{\partial^2 f}{\partial y\,\partial x}$ and $\frac{\partial^2 f}{\partial x\,\partial y}$ for the function $f(x,y) = e^{x^2 y} + xy^3$.

Solution: Proceeding as before, we have

$$\frac{\partial f}{\partial x} = 2xye^{x^2y} + y^3 \qquad \frac{\partial f}{\partial y} = x^2e^{x^2y} + 3xy^2$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(2xye^{x^2y} + y^3) \qquad \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(x^2e^{x^2y} + 3xy^2)$$

$$= 2ye^{x^2y} + 4x^2y^2e^{x^2y} \qquad = x^4e^{x^2y} + 6xy$$

$$\frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial}{\partial y}(2xye^{x^2y} + y^3) \qquad \frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial}{\partial x}(x^2e^{x^2y} + 3xy^2)$$

$$= 2xe^{x^2y} + 2x^3ye^{x^2y} + 3y^2 \qquad = 2xe^{x^2y} + 2x^3ye^{x^2y} + 3y^2$$

Higher-order partial derivatives that are taken with respect to different variables, such as $\frac{\partial^2 f}{\partial y\,\partial x}$ and $\frac{\partial^2 f}{\partial x\,\partial y}$, are called **mixed partial derivatives**. Notice in the above example that $\frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial^2 f}{\partial x\,\partial y}$. It turns that this will usually be the case. Specifically, whenever both $\frac{\partial^2 f}{\partial y\,\partial x}$ and $\frac{\partial^2 f}{\partial x\,\partial y}$ are continuous at a point (a,b), then they are equal at that point.[2] All the functions we will deal with will have continuous partial derivatives of all orders, so you can assume in the remainder of the text that

$$\frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial^2 f}{\partial x\,\partial y} \quad \text{for all } (x,y) \text{ in the domain of } f.$$

In other words, it doesn't matter in which order you take partial derivatives. This applies even to mixed partial derivatives of order 3 or higher.

The notation for partial derivatives varies. All of the following are equivalent:

$$\frac{\partial f}{\partial x} : \; f_x(x,y)\,, \quad f_1(x,y)\,, \quad D_x(x,y)\,, \quad D_1(x,y)$$

$$\frac{\partial f}{\partial y} : \; f_y(x,y)\,, \quad f_2(x,y)\,, \quad D_y(x,y)\,, \quad D_2(x,y)$$

$$\frac{\partial^2 f}{\partial x^2} : \; f_{xx}(x,y)\,, \quad f_{11}(x,y)\,, \quad D_{xx}(x,y)\,, \quad D_{11}(x,y)$$

$$\frac{\partial^2 f}{\partial y^2} : \; f_{yy}(x,y)\,, \quad f_{22}(x,y)\,, \quad D_{yy}(x,y)\,, \quad D_{22}(x,y)$$

$$\frac{\partial^2 f}{\partial y\,\partial x} : \; f_{xy}(x,y)\,, \quad f_{12}(x,y)\,, \quad D_{xy}(x,y)\,, \quad D_{12}(x,y)$$

$$\frac{\partial^2 f}{\partial x\,\partial y} : \; f_{yx}(x,y)\,, \quad f_{21}(x,y)\,, \quad D_{yx}(x,y)\,, \quad D_{21}(x,y)$$

[2] See pp. 214-216 in TAYLOR and MANN for a proof.

Exercises

A

For Exercises 1-16, find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

1. $f(x,y) = x^2 + y^2$

2. $f(x,y) = \cos(x+y)$

3. $f(x,y) = \sqrt{x^2+y+4}$

4. $f(x,y) = \dfrac{x+1}{y+1}$

5. $f(x,y) = e^{xy} + xy$

6. $f(x,y) = x^2 - y^2 + 6xy + 4x - 8y + 2$

7. $f(x,y) = x^4$

8. $f(x,y) = x + 2y$

9. $f(x,y) = \sqrt{x^2+y^2}$

10. $f(x,y) = \sin(x+y)$

11. $f(x,y) = \sqrt[3]{x^2+y+4}$

12. $f(x,y) = \dfrac{xy+1}{x+y}$

13. $f(x,y) = e^{-(x^2+y^2)}$

14. $f(x,y) = \ln(xy)$

15. $f(x,y) = \sin(xy)$

16. $f(x,y) = \tan(x+y)$

For Exercises 17-26, find $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial y^2}$ and $\frac{\partial^2 f}{\partial y \partial x}$ (use Exercises 1-8, 14, 15).

17. $f(x,y) = x^2 + y^2$

18. $f(x,y) = \cos(x+y)$

19. $f(x,y) = \sqrt{x^2+y+4}$

20. $f(x,y) = \dfrac{x+1}{y+1}$

21. $f(x,y) = e^{xy} + xy$

22. $f(x,y) = x^2 - y^2 + 6xy + 4x - 8y + 2$

23. $f(x,y) = x^4$

24. $f(x,y) = x + 2y$

25. $f(x,y) = \ln(xy)$

26. $f(x,y) = \sin(xy)$

B

27. Show that the function $f(x,y) = \sin(x+y) + \cos(x-y)$ satisfies the *wave equation*

$$\frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial y^2} = 0\ .$$

The wave equation is an example of a *partial differential equation*.

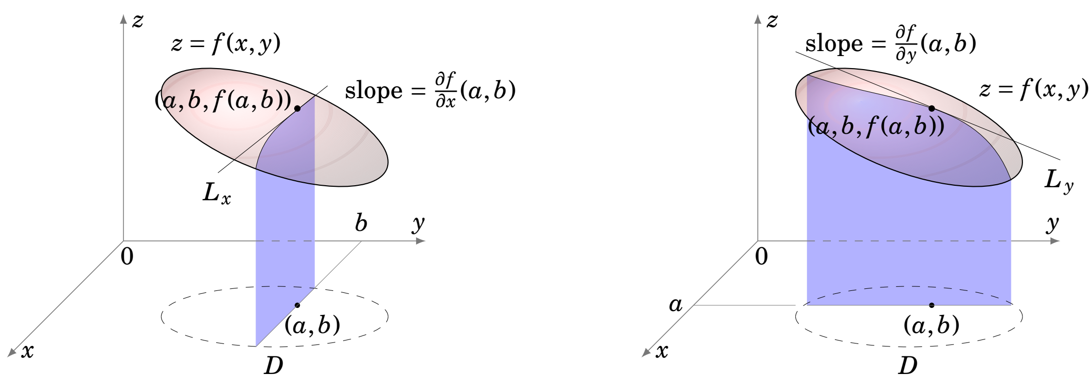
28. Let u and v be twice-differentiable functions of a single variable, and let $c \neq 0$ be a constant. Show that $f(x,y) = u(x+cy) + v(x-cy)$ is a solution of the *general one-dimensional wave equation*[3]

$$\frac{\partial^2 f}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 f}{\partial y^2} = 0\ .$$

[3] Conversely, it turns out that *any* solution must be of this form. See Ch. 1 in WEINBERGER.

2.3 Tangent Plane to a Surface

In the previous section we mentioned that the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ can be thought of as the rate of change of a function $z = f(x,y)$ in the positive x and y directions, respectively. Recall that the derivative $\frac{dy}{dx}$ of a function $y = f(x)$ has a geometric meaning, namely as the slope of the tangent line to the graph of f at the point $(x, f(x))$ in $\mathbb{R}^2$. There is a similar geometric meaning to the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ of a function $z = f(x,y)$: given a point (a,b) in the domain D of $f(x,y)$, the trace of the surface described by $z = f(x,y)$ in the plane $y = b$ is a curve in $\mathbb{R}^3$ through the point $(a,b,f(a,b))$, and the slope of the tangent line L_x to that curve at that point is $\frac{\partial f}{\partial x}(a,b)$. Similarly, $\frac{\partial f}{\partial y}(a,b)$ is the slope of the tangent line L_y to the trace of the surface $z = f(x,y)$ in the plane $x = a$ (see Figure 2.3.1).



(a) Tangent line L_x in the plane $y = b$ **(b)** Tangent line L_y in the plane $x = a$

Figure 2.3.1 Partial derivatives as slopes

Since the derivative $\frac{dy}{dx}$ of a function $y = f(x)$ is used to find the tangent line to the graph of f (which is a curve in $\mathbb{R}^2$), you might expect that partial derivatives can be used to define a *tangent plane* to the graph of a surface $z = f(x,y)$. This indeed turns out to be the case. First, we need a definition of a tangent plane. The intuitive idea is that a tangent plane "just touches" a surface at a point. The formal definition mimics the intuitive notion of a tangent line to a curve.

Definition 2.4. Let $z = f(x,y)$ be the equation of a surface S in $\mathbb{R}^3$, and let $P = (a,b,c)$ be a point on S. Let T be a plane which contains the point P, and let $Q = (x,y,z)$ represent a generic point on the surface S. If the (acute) angle between the vector $\overrightarrow{PQ}$ and the plane T approaches zero as the point Q approaches P along the surface S, then we call T the **tangent plane** to S at P.

Note that since two lines in $\mathbb{R}^3$ determine a plane, then the two tangent lines to the surface $z = f(x,y)$ in the x and y directions described in Figure 2.3.1 are contained in the tangent plane at that point, *if the tangent plane exists at that point*. The existence of those two

tangent lines does not by itself guarantee the existence of the tangent plane. It is possible that if we take the trace of the surface in the plane $x - y = 0$ (which makes a 45° angle with the positive x-axis), the resulting curve in that plane may have a tangent line which is not in the plane determined by the other two tangent lines, or it may not have a tangent line at all at that point. Luckily, it turns out[4] that if $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist in a region around a point (a,b) and are continuous at (a,b) then the tangent plane to the surface $z = f(x,y)$ will exist at the point $(a,b,f(a,b))$. In this text, those conditions will always hold.

Suppose that we want an equation of the tangent plane T to the surface $z = f(x,y)$ at a point $(a,b,f(a,b))$. Let L_x and L_y be the tangent lines to the traces of the surface in the planes $y = b$ and $x = a$, respectively (as in Figure 2.3.2), and suppose that the conditions for T to exist do hold. Then the equation for T is

$$A(x-a) + B(y-b) + C(z - f(a,b)) = 0 \tag{2.4}$$

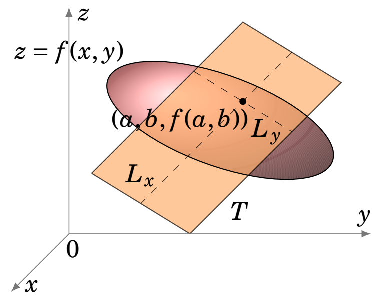


Figure 2.3.2 Tangent plane

where $\mathbf{n} = (A,B,C)$ is a normal vector to the plane T. Since T contains the lines L_x and L_y, then all we need are vectors $\mathbf{v}_x$ and $\mathbf{v}_y$ that are parallel to L_x and L_y, respectively, and then let $\mathbf{n} = \mathbf{v}_x \times \mathbf{v}_y$.

Since the slope of L_x is $\frac{\partial f}{\partial x}(a,b)$, then the vector $\mathbf{v}_x = (1,0,\frac{\partial f}{\partial x}(a,b))$ is parallel to L_x (since $\mathbf{v}_x$ lies in the xz-plane and lies in a line with slope $\frac{\frac{\partial f}{\partial x}(a,b)}{1} = \frac{\partial f}{\partial x}(a,b)$. See Figure 2.3.3). Similarly, the vector $\mathbf{v}_y = (0,1,\frac{\partial f}{\partial y}(a,b))$ is parallel to L_y. Hence, the vector

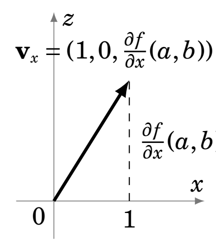


Figure 2.3.3

$$\mathbf{n} = \mathbf{v}_x \times \mathbf{v}_y = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 0 & \frac{\partial f}{\partial x}(a,b) \\ 0 & 1 & \frac{\partial f}{\partial y}(a,b) \end{vmatrix} = -\frac{\partial f}{\partial x}(a,b)\,\mathbf{i} - \frac{\partial f}{\partial y}(a,b)\,\mathbf{j} + \mathbf{k}$$

is normal to the plane T. Thus the equation of T is

$$-\frac{\partial f}{\partial x}(a,b)(x-a) - \frac{\partial f}{\partial y}(a,b)(y-b) + z - f(a,b) = 0\ . \tag{2.5}$$

Multiplying both sides by -1, we have the following result:

The equation of the tangent plane to the surface $z = f(x,y)$ at the point $(a,b,f(a,b))$ is

$$\frac{\partial f}{\partial x}(a,b)(x-a) + \frac{\partial f}{\partial y}(a,b)(y-b) - z + f(a,b) = 0 \tag{2.6}$$

[4]See TAYLOR and MANN, § 6.4.

Example 2.13. Find the equation of the tangent plane to the surface $z = x^2 + y^2$ at the point $(1,2,5)$.

Solution: For the function $f(x,y) = x^2 + y^2$, we have $\frac{\partial f}{\partial x} = 2x$ and $\frac{\partial f}{\partial y} = 2y$, so the equation of the tangent plane at the point $(1,2,5)$ is

$$2(1)(x-1) + 2(2)(y-2) - z + 5 = 0 \text{ , or}$$
$$2x + 4y - z - 5 = 0 \text{ .}$$

In a similar fashion, it can be shown that if a surface is defined implicitly by an equation of the form $F(x,y,z) = 0$, then the tangent plane to the surface at a point (a,b,c) is given by the equation

$$\frac{\partial F}{\partial x}(a,b,c)(x-a) + \frac{\partial F}{\partial y}(a,b,c)(y-b) + \frac{\partial F}{\partial z}(a,b,c)(z-c) = 0 \text{ .} \tag{2.7}$$

Note that formula (2.6) is the special case of formula (2.7) where $F(x,y,z) = f(x,y) - z$.

Example 2.14. Find the equation of the tangent plane to the surface $x^2 + y^2 + z^2 = 9$ at the point $(2,2,-1)$.

Solution: For the function $F(x,y,z) = x^2 + y^2 + z^2 - 9$, we have $\frac{\partial F}{\partial x} = 2x$, $\frac{\partial F}{\partial y} = 2y$, and $\frac{\partial F}{\partial z} = 2z$, so the equation of the tangent plane at $(2,2,-1)$ is

$$2(2)(x-2) + 2(2)(y-2) + 2(-1)(z+1) = 0 \text{ , or}$$
$$2x + 2y - z - 9 = 0 \text{ .}$$

Exercises

A

For Exercises 1-6, find the equation of the tangent plane to the surface $z = f(x,y)$ at the point P.

1. $f(x,y) = x^2 + y^3$, $P = (1,1,2)$

2. $f(x,y) = xy$, $P = (1,-1,-1)$

3. $f(x,y) = x^2 y$, $P = (-1,1,1)$

4. $f(x,y) = xe^y$, $P = (1,0,1)$

5. $f(x,y) = x + 2y$, $P = (2,1,4)$

6. $f(x,y) = \sqrt{x^2 + y^2}$, $P = (3,4,5)$

For Exercises 7-10, find the equation of the tangent plane to the given surface at the point P.

7. $\frac{x^2}{4} + \frac{y^2}{9} + \frac{z^2}{16} = 1$, $P = \left(1, 2, \frac{2\sqrt{11}}{3}\right)$

8. $x^2 + y^2 + z^2 = 9$, $P = (0,0,3)$

9. $x^2 + y^2 - z^2 = 0$, $P = (3,4,5)$

10. $x^2 + y^2 = 4$, $P = (\sqrt{3},1,0)$

2.4 Directional Derivatives and the Gradient

For a function $z = f(x,y)$, we learned that the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ represent the (instantaneous) rate of change of f in the positive x and y directions, respectively. What about other directions? It turns out that we can find the rate of change in *any* direction using a more general type of derivative called a *directional derivative*.

Definition 2.5. Let $f(x,y)$ be a real-valued function with domain D in $\mathbb{R}^2$, and let (a,b) be a point in D. Let $\mathbf{v}$ be a unit vector in $\mathbb{R}^2$. Then the **directional derivative of f at (a,b) in the direction of $\mathbf{v}$**, denoted by $D_{\mathbf{v}} f(a,b)$, is defined as

$$D_{\mathbf{v}} f(a,b) = \lim_{h\to 0} \frac{f((a,b)+h\mathbf{v}) - f(a,b)}{h} \tag{2.8}$$

Notice in the definition that we seem to be treating the point (a,b) as a vector, since we are adding the vector $h\mathbf{v}$ to it. But this is just the usual idea of identifying vectors with their terminal points, which the reader should be used to by now. If we were to write the vector $\mathbf{v}$ as $\mathbf{v} = (v_1, v_2)$, then

$$D_{\mathbf{v}} f(a,b) = \lim_{h\to 0} \frac{f(a+hv_1, b+hv_2) - f(a,b)}{h} \;. \tag{2.9}$$

From this we can immediately recognize that the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are special cases of the directional derivative with $\mathbf{v} = \mathbf{i} = (1,0)$ and $\mathbf{v} = \mathbf{j} = (0,1)$, respectively. That is, $\frac{\partial f}{\partial x} = D_{\mathbf{i}} f$ and $\frac{\partial f}{\partial y} = D_{\mathbf{j}} f$. Since there are many vectors with the same direction, we use a unit vector in the definition, as that represents a "standard" vector for a given direction.

If $f(x,y)$ has continuous partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ (which will always be the case in this text), then there is a simple formula for the directional derivative:

Theorem 2.2. Let $f(x,y)$ be a real-valued function with domain D in $\mathbb{R}^2$ such that the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist and are continuous in D. Let (a,b) be a point in D, and let $\mathbf{v} = (v_1, v_2)$ be a unit vector in $\mathbb{R}^2$. Then

$$D_{\mathbf{v}} f(a,b) = v_1 \frac{\partial f}{\partial x}(a,b) + v_2 \frac{\partial f}{\partial y}(a,b) \;. \tag{2.10}$$

Proof: Note that if $\mathbf{v} = \mathbf{i} = (1,0)$ then the above formula reduces to $D_{\mathbf{v}} f(a,b) = \frac{\partial f}{\partial x}(a,b)$, which we know is true since $D_{\mathbf{i}} f = \frac{\partial f}{\partial x}$, as we noted earlier. Similarly, for $\mathbf{v} = \mathbf{j} = (0,1)$ the formula reduces to $D_{\mathbf{v}} f(a,b) = \frac{\partial f}{\partial y}(a,b)$, which is true since $D_{\mathbf{j}} f = \frac{\partial f}{\partial y}$. So since $\mathbf{i} = (1,0)$ and $\mathbf{j} = (0,1)$ are the only unit vectors in $\mathbb{R}^2$ with a zero component, then we need only show the formula holds for unit vectors $\mathbf{v} = (v_1, v_2)$ with $v_1 \neq 0$ and $v_2 \neq 0$. So fix such a vector $\mathbf{v}$ and fix a number $h \neq 0$.

Then

$$f(a+hv_1,b+hv_2)-f(a,b) = f(a+hv_1,b+hv_2)-f(a+hv_1,b)+f(a+hv_1,b)-f(a,b)\;. \quad (2.11)$$

Since $h \neq 0$ and $v_2 \neq 0$, then $hv_2 \neq 0$ and thus any number c between b and $b+hv_2$ can be written as $c = b+\alpha hv_2$ for some number $0 < \alpha < 1$. So since the function $f(a+hv_1, y)$ is a real-valued function of y (since $a+hv_1$ is a fixed number), then the Mean Value Theorem from single-variable calculus can be applied to the function $g(y) = f(a+hv_1, y)$ on the interval $[b, b+hv_2]$ (or $[b+hv_2, b]$ if one of h or v_2 is negative) to find a number $0 < \alpha < 1$ such that

$$\frac{\partial f}{\partial y}(a+hv_1,b+\alpha hv_2) = g'(b+\alpha hv_2) = \frac{g(b+hv_2)-g(b)}{b+hv_2-b} = \frac{f(a+hv_1,b+hv_2)-f(a+hv_1,b)}{hv_2}$$

and so

$$f(a+hv_1,b+hv_2)-f(a+hv_1,b) \;=\; hv_2\frac{\partial f}{\partial y}(a+hv_1,b+\alpha hv_2)\;.$$

By a similar argument, there exists a number $0 < \beta < 1$ such that

$$f(a+hv_1,b)-f(a,b) \;=\; hv_1\frac{\partial f}{\partial x}(a+\beta hv_1,b)\;.$$

Thus, by equation (2.11), we have

$$\begin{aligned}
\frac{f(a+hv_1,b+hv_2)-f(a,b)}{h} &= \frac{hv_2\frac{\partial f}{\partial y}(a+hv_1,b+\alpha hv_2)+hv_1\frac{\partial f}{\partial x}(a+\beta hv_1,b)}{h}\\
&= v_2\frac{\partial f}{\partial y}(a+hv_1,b+\alpha hv_2)+v_1\frac{\partial f}{\partial x}(a+\beta hv_1,b)
\end{aligned}$$

so by formula (2.9) we have

$$\begin{aligned}
D_{\mathbf{v}}f(a,b) &= \lim_{h\to 0}\frac{f(a+hv_1,b+hv_2)-f(a,b)}{h}\\
&= \lim_{h\to 0}\left[v_2\frac{\partial f}{\partial y}(a+hv_1,b+\alpha hv_2)+v_1\frac{\partial f}{\partial x}(a+\beta hv_1,b)\right]\\
&= v_2\frac{\partial f}{\partial y}(a,b)+v_1\frac{\partial f}{\partial x}(a,b) \quad \text{by the continuity of } \frac{\partial f}{\partial x} \text{ and } \frac{\partial f}{\partial y}\text{, so}\\
D_{\mathbf{v}}f(a,b) &= v_1\frac{\partial f}{\partial x}(a,b)+v_2\frac{\partial f}{\partial y}(a,b)
\end{aligned}$$

after reversing the order of summation. **QED**

Note that $D_{\mathbf{v}}f(a,b) = \mathbf{v}\cdot\left(\frac{\partial f}{\partial x}(a,b), \frac{\partial f}{\partial y}(a,b)\right)$. The second vector has a special name:

Definition 2.6. For a real-valued function $f(x,y)$, the **gradient** of f, denoted by ∇f, is the vector

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \tag{2.12}$$

in $\mathbb{R}^2$. For a real-valued function $f(x,y,z)$, the gradient is the vector

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) \tag{2.13}$$

in $\mathbb{R}^3$. The symbol ∇ is pronounced "del".[5]

Corollary 2.3. $D_{\mathbf{v}} f = \mathbf{v} \cdot \nabla f$

Example 2.15. Find the directional derivative of $f(x,y) = xy^2 + x^3 y$ at the point $(1,2)$ in the direction of $\mathbf{v} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

Solution: We see that $\nabla f = (y^2 + 3x^2 y, 2xy + x^3)$, so

$$D_{\mathbf{v}} f(1,2) \;=\; \mathbf{v} \cdot \nabla f(1,2) \;=\; \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}\right) \cdot (2^2 + 3(1)^2(2), 2(1)(2) + 1^3) \;=\; \tfrac{15}{\sqrt{2}}$$

A real-valued function $z = f(x,y)$ whose partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist and are continuous is called *continuously differentiable*. Assume that $f(x,y)$ is such a function and that $\nabla f \neq \mathbf{0}$. Let c be a real number in the range of f and let $\mathbf{v}$ be a unit vector in $\mathbb{R}^2$ which is tangent to the level curve $f(x,y) = c$ (see Figure 2.4.1).

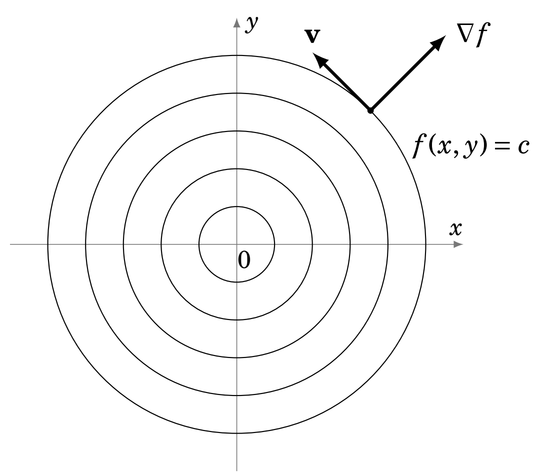


Figure 2.4.1

[5]Sometimes the notation grad(f) is used instead of ∇f.

The value of $f(x,y)$ is constant along a level curve, so since $\mathbf{v}$ is a tangent vector to this curve, then the rate of change of f in the direction of $\mathbf{v}$ is 0, i.e. $D_{\mathbf{v}}f = 0$. But we know that $D_{\mathbf{v}}f = \mathbf{v}\cdot\nabla f = \|\mathbf{v}\|\,\|\nabla f\|\cos\theta$, where θ is the angle between $\mathbf{v}$ and ∇f. So since $\|\mathbf{v}\| = 1$ then $D_{\mathbf{v}}f = \|\nabla f\|\cos\theta$. So since $\nabla f \neq \mathbf{0}$ then $D_{\mathbf{v}}f = 0 \Rightarrow \cos\theta = 0 \Rightarrow \theta = 90°$. In other words, $\nabla f \perp \mathbf{v}$, which means that ∇f is *normal* to the level curve.

In general, for any unit vector $\mathbf{v}$ in $\mathbb{R}^2$, we still have $D_{\mathbf{v}}f = \|\nabla f\|\cos\theta$, where θ is the angle between $\mathbf{v}$ and ∇f. At a fixed point (x,y) the length $\|\nabla f\|$ is fixed, and the value of $D_{\mathbf{v}}f$ then varies as θ varies. The largest value that $D_{\mathbf{v}}f$ can take is when $\cos\theta = 1$ ($\theta = 0°$), while the smallest value occurs when $\cos\theta = -1$ ($\theta = 180°$). In other words, the value of the function f increases the fastest in the direction of ∇f (since $\theta = 0°$ in that case), and the value of f decreases the fastest in the direction of $-\nabla f$ (since $\theta = 180°$ in that case). We have thus proved the following theorem:

Theorem 2.4. Let $f(x,y)$ be a continuously differentiable real-valued function, with $\nabla f \neq \mathbf{0}$. Then:

(a) The gradient ∇f is normal to any level curve $f(x,y) = c$.

(b) The value of $f(x,y)$ increases the fastest in the direction of ∇f.

(c) The value of $f(x,y)$ decreases the fastest in the direction of $-\nabla f$.

Example 2.16. In which direction does the function $f(x,y) = xy^2 + x^3y$ increase the fastest from the point $(1,2)$? In which direction does it decrease the fastest?

Solution: Since $\nabla f = (y^2 + 3x^2y, 2xy + x^3)$, then $\nabla f(1,2) = (10,5) \neq \mathbf{0}$. A unit vector in that direction is $\mathbf{v} = \frac{\nabla f}{\|\nabla f\|} = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)$. Thus, f increases the fastest in the direction of $\left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)$ and decreases the fastest in the direction of $\left(\frac{-2}{\sqrt{5}}, \frac{-1}{\sqrt{5}}\right)$.

Though we proved Theorem 2.4 for functions of two variables, a similar argument can be used to show that it also applies to functions of three or more variables. Likewise, the directional derivative in the three-dimensional case can also be defined by the formula $D_{\mathbf{v}}f = \mathbf{v}\cdot\nabla f$.

Example 2.17. The temperature T of a solid is given by the function $T(x,y,z) = e^{-x} + e^{-2y} + e^{4z}$, where x, y, z are space coordinates relative to the center of the solid. In which direction from the point $(1,1,1)$ will the temperature decrease the fastest?

Solution: Since $\nabla f = (-e^{-x}, -2e^{-2y}, 4e^{4z})$, then the temperature will decrease the fastest in the direction of $-\nabla f(1,1,1) = (e^{-1}, 2e^{-2}, -4e^{4})$.

Exercises

A

For Exercises 1-10, compute the gradient ∇f.

1. $f(x,y) = x^2 + y^2 - 1$

2. $f(x,y) = \dfrac{1}{x^2 + y^2}$

3. $f(x,y) = \sqrt{x^2 + y^2 + 4}$

4. $f(x,y) = x^2 e^y$

5. $f(x,y) = \ln(xy)$

6. $f(x,y) = 2x + 5y$

7. $f(x,y,z) = \sin(xyz)$

8. $f(x,y,z) = x^2 e^{yz}$

9. $f(x,y,z) = x^2 + y^2 + z^2$

10. $f(x,y,z) = \sqrt{x^2 + y^2 + z^2}$

For Exercises 11-14, find the directional derivative of f at the point P in the direction of $\mathbf{v} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

11. $f(x,y) = x^2 + y^2 - 1$, $P = (1,1)$

12. $f(x,y) = \dfrac{1}{x^2 + y^2}$, $P = (1,1)$

13. $f(x,y) = \sqrt{x^2 + y^2 + 4}$, $P = (1,1)$

14. $f(x,y) = x^2 e^y$, $P = (1,1)$

For Exercises 15-16, find the directional derivative of f at the point P in the direction of $\mathbf{v} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$.

15. $f(x,y,z) = \sin(xyz)$, $P = (1,1,1)$

16. $f(x,y,z) = x^2 e^{yz}$, $P = (1,1,1)$

17. Repeat Example 2.16 at the point $(2,3)$.

18. Repeat Example 2.17 at the point $(3,1,2)$.

B

For Exercises 19-26, let $f(x,y)$ and $g(x,y)$ be continuously differentiable real-valued functions, let c be a constant, and let $\mathbf{v}$ be a unit vector in $\mathbb{R}^2$. Show that:

19. $\nabla(cf) = c\nabla f$

20. $\nabla(f+g) = \nabla f + \nabla g$

21. $\nabla(fg) = f\nabla g + g\nabla f$

22. $\nabla(f/g) = \dfrac{g\nabla f - f\nabla g}{g^2}$ if $g(x,y) \neq 0$

23. $D_{-\mathbf{v}} f = -D_{\mathbf{v}} f$

24. $D_{\mathbf{v}}(cf) = c D_{\mathbf{v}} f$

25. $D_{\mathbf{v}}(f+g) = D_{\mathbf{v}} f \; + \; D_{\mathbf{v}} g$

26. $D_{\mathbf{v}}(fg) = f D_{\mathbf{v}} g \; + \; g D_{\mathbf{v}} f$

27. The function $r(x,y) = \sqrt{x^2 + y^2}$ is the length of the position vector $\mathbf{r} = x\,\mathbf{i} + y\,\mathbf{j}$ for each point (x,y) in $\mathbb{R}^2$. Show that $\nabla r = \dfrac{1}{r}\mathbf{r}$ when $(x,y) \neq (0,0)$, and that $\nabla(r^2) = 2\mathbf{r}$.

2.5 Maxima and Minima

The gradient can be used to find *extreme points* of real-valued functions of several variables, that is, points where the function has a *local maximum* or *local minimum*. We will consider only functions of two variables; functions of three or more variables require methods using linear algebra.

Definition 2.7. Let $f(x,y)$ be a real-valued function, and let (a,b) be a point in the domain of f. We say that f has a **local maximum** at (a,b) if $f(x,y) \le f(a,b)$ for all (x,y) inside some disk of positive radius centered at (a,b), i.e. there is some sufficiently small $r > 0$ such that $f(x,y) \le f(a,b)$ for all (x,y) for which $(x-a)^2 + (y-b)^2 < r^2$.

Likewise, we say that f has a **local minimum** at (a,b) if $f(x,y) \ge f(a,b)$ for all (x,y) inside some disk of positive radius centered at (a,b).

If $f(x,y) \le f(a,b)$ for all (x,y) in the domain of f, then f has a **global maximum** at (a,b). If $f(x,y) \ge f(a,b)$ for all (x,y) in the domain of f, then f has a **global minimum** at (a,b).

Suppose that (a,b) is a local maximum point for $f(x,y)$, and that the first-order partial derivatives of f exist at (a,b). We know that $f(a,b)$ is the largest value of $f(x,y)$ as (x,y) goes in all directions from the point (a,b), in some sufficiently small disk centered at (a,b). In particular, $f(a,b)$ is the largest value of f in the x direction (around the point (a,b)), that is, the single-variable function $g(x) = f(x,b)$ has a local maximum at $x = a$. So we know that $g'(a) = 0$. Since $g'(x) = \frac{\partial f}{\partial x}(x,b)$, then $\frac{\partial f}{\partial x}(a,b) = 0$. Similarly, $f(a,b)$ is the largest value of f near (a,b) in the y direction and so $\frac{\partial f}{\partial y}(a,b) = 0$. We thus have the following theorem:

Theorem 2.5. Let $f(x,y)$ be a real-valued function such that both $\frac{\partial f}{\partial x}(a,b)$ and $\frac{\partial f}{\partial y}(a,b)$ exist. Then a necessary condition for $f(x,y)$ to have a local maximum or minimum at (a,b) is that $\nabla f(a,b) = \mathbf{0}$.

Note: Theorem 2.5 can be extended to apply to functions of three or more variables.

A point (a,b) where $\nabla f(a,b) = \mathbf{0}$ is called a **critical point** for the function $f(x,y)$. So given a function $f(x,y)$, to find the critical points of f you have to solve the equations $\frac{\partial f}{\partial x}(x,y) = 0$ and $\frac{\partial f}{\partial y}(x,y) = 0$ simultaneously for (x,y). Similar to the single-variable case, the *necessary* condition that $\nabla f(a,b) = \mathbf{0}$ is not always *sufficient* to guarantee that a critical point is a local maximum or minimum.

Example 2.18. The function $f(x,y) = xy$ has a critical point at $(0,0)$: $\frac{\partial f}{\partial x} = y = 0 \Rightarrow y = 0$, and $\frac{\partial f}{\partial y} = x = 0 \Rightarrow x = 0$, so $(0,0)$ is the only critical point. But clearly f does not have a local maximum or minimum at $(0,0)$ since any disk around $(0,0)$ contains points (x,y) where the values of x and y have the same sign (so that $f(x,y) = xy > 0 = f(0,0)$) and different signs (so that $f(x,y) = xy < 0 = f(0,0)$). In fact, along the path $y = x$ in $\mathbb{R}^2$, $f(x,y) = x^2$, which has a

local minimum at $(0,0)$, while along the path $y = -x$ we have $f(x, y) = -x^2$, which has a local maximum at $(0,0)$. So $(0,0)$ is an example of a *saddle point*, i.e. it is a local maximum in one direction and a local minimum in another direction. The graph of $f(x, y)$ is shown in Figure 2.5.1, which is a hyperbolic paraboloid.

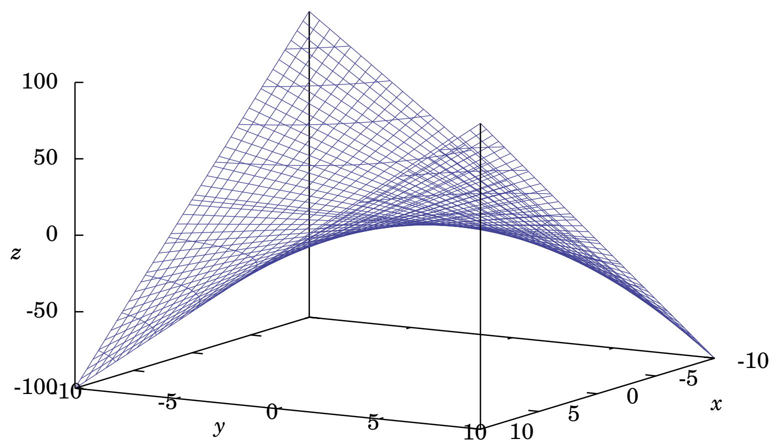


Figure 2.5.1 $f(x, y) = xy$, saddle point at $(0,0)$

The following theorem gives sufficient conditions for a critical point to be a local maximum or minimum of a *smooth function* (i.e. a function whose partial derivatives of all orders exist and are continuous), which we will not prove here.[6]

Theorem 2.6. Let $f(x, y)$ be a smooth real-valued function, with a critical point at (a, b) (i.e. $\nabla f(a, b) = \mathbf{0}$). Define

$$D = \frac{\partial^2 f}{\partial x^2}(a,b)\,\frac{\partial^2 f}{\partial y^2}(a,b) - \left(\frac{\partial^2 f}{\partial y\,\partial x}(a,b)\right)^2$$

Then

(a) if $D > 0$ and $\frac{\partial^2 f}{\partial x^2}(a,b) > 0$, then f has a local minimum at (a, b)

(b) if $D > 0$ and $\frac{\partial^2 f}{\partial x^2}(a,b) < 0$, then f has a local maximum at (a, b)

(c) if $D < 0$, then f has neither a local minimum nor a local maximum at (a, b)

(d) if $D = 0$, then the test fails.

[6] See TAYLOR and MANN, § 7.6.

If condition (c) holds, then (a,b) is a *saddle point*. Note that the assumption that $f(x,y)$ is smooth means that

$$D = \begin{vmatrix} \dfrac{\partial^2 f}{\partial x^2}(a,b) & \dfrac{\partial^2 f}{\partial y\,\partial x}(a,b) \\ \dfrac{\partial^2 f}{\partial x\,\partial y}(a,b) & \dfrac{\partial^2 f}{\partial y^2}(a,b) \end{vmatrix}$$

since $\frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial^2 f}{\partial x\,\partial y}$. Also, if $D > 0$ then $\frac{\partial^2 f}{\partial x^2}(a,b)\frac{\partial^2 f}{\partial y^2}(a,b) = D + \left(\frac{\partial^2 f}{\partial y\,\partial x}(a,b)\right)^2 > 0$, and so $\frac{\partial^2 f}{\partial x^2}(a,b)$ and $\frac{\partial^2 f}{\partial y^2}(a,b)$ have the same sign. This means that in parts (a) and (b) of the theorem one can replace $\frac{\partial^2 f}{\partial x^2}(a,b)$ by $\frac{\partial^2 f}{\partial y^2}(a,b)$ if desired.

Example 2.19. Find all local maxima and minima of $f(x,y) = x^2 + xy + y^2 - 3x$.

Solution: First find the critical points, i.e. where $\nabla f = \mathbf{0}$. Since

$$\frac{\partial f}{\partial x} = 2x + y - 3 \quad \text{and} \quad \frac{\partial f}{\partial y} = x + 2y$$

then the critical points (x,y) are the common solutions of the equations

$$\begin{aligned} 2x + y - 3 &= 0 \\ x + 2y \phantom{{}-3} &= 0 \end{aligned}$$

which has the unique solution $(x,y) = (2,-1)$. So $(2,-1)$ is the only critical point.

To use Theorem 2.6, we need the second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = 2\ , \quad \frac{\partial^2 f}{\partial y^2} = 2\ , \quad \frac{\partial^2 f}{\partial y\,\partial x} = 1$$

and so

$$D \;=\; \frac{\partial^2 f}{\partial x^2}(2,-1)\frac{\partial^2 f}{\partial y^2}(2,-1) - \left(\frac{\partial^2 f}{\partial y\,\partial x}(2,-1)\right)^2 \;=\; (2)(2) - 1^2 \;=\; 3 \;>\; 0$$

and $\frac{\partial^2 f}{\partial x^2}(2,-1) = 2 > 0$. Thus, $(2,-1)$ is a local minimum.

Example 2.20. Find all local maxima and minima of $f(x,y) = xy - x^3 - y^2$.

Solution: First find the critical points, i.e. where $\nabla f = \mathbf{0}$. Since

$$\frac{\partial f}{\partial x} = y - 3x^2 \quad \text{and} \quad \frac{\partial f}{\partial y} = x - 2y$$

then the critical points (x, y) are the common solutions of the equations

$$y - 3x^2 = 0$$
$$x - 2y = 0$$

The first equation yields $y = 3x^2$, substituting that into the second equation yields $x - 6x^2 = 0$, which has the solutions $x = 0$ and $x = \frac{1}{6}$. So $x = 0 \Rightarrow y = 3(0) = 0$ and $x = \frac{1}{6} \Rightarrow y = 3\left(\frac{1}{6}\right)^2 = \frac{1}{12}$. So the critical points are $(x, y) = (0, 0)$ and $(x, y) = \left(\frac{1}{6}, \frac{1}{12}\right)$.

To use Theorem 2.6, we need the second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = -6x\ , \quad \frac{\partial^2 f}{\partial y^2} = -2\ , \quad \frac{\partial^2 f}{\partial y\,\partial x} = 1$$

So

$$D \ = \ \frac{\partial^2 f}{\partial x^2}(0,0)\,\frac{\partial^2 f}{\partial y^2}(0,0) - \left(\frac{\partial^2 f}{\partial y\,\partial x}(0,0)\right)^2 \ = \ (-6(0))(-2) - 1^2 \ = \ -1 \ < \ 0$$

and thus $(0, 0)$ is a saddle point. Also,

$$D \ = \ \frac{\partial^2 f}{\partial x^2}\left(\tfrac{1}{6}, \tfrac{1}{12}\right)\frac{\partial^2 f}{\partial y^2}\left(\tfrac{1}{6}, \tfrac{1}{12}\right) - \left(\frac{\partial^2 f}{\partial y\,\partial x}\left(\tfrac{1}{6}, \tfrac{1}{12}\right)\right)^2 \ = \ (-6(\frac{1}{6}))(-2) - 1^2 \ = \ 1 \ > \ 0$$

and $\frac{\partial^2 f}{\partial x^2}\left(\frac{1}{6}, \frac{1}{12}\right) = -1 < 0$. Thus, $\left(\frac{1}{6}, \frac{1}{12}\right)$ is a local maximum.

Example 2.21. Find all local maxima and minima of $f(x, y) = (x-2)^4 + (x-2y)^2$.

Solution: First find the critical points, i.e. where $\nabla f = \mathbf{0}$. Since

$$\frac{\partial f}{\partial x} = 4(x-2)^3 + 2(x-2y) \quad \text{and} \quad \frac{\partial f}{\partial y} = -4(x-2y)$$

then the critical points (x, y) are the common solutions of the equations

$$4(x-2)^3 + 2(x-2y) = 0$$
$$-4(x-2y) = 0$$

The second equation yields $x = 2y$, substituting that into the first equation yields $4(2y-2)^3 = 0$, which has the solution $y = 1$, and so $x = 2(1) = 2$. Thus, $(2, 1)$ is the only critical point.

To use Theorem 2.6, we need the second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = 12(x-2)^2 + 2\ , \quad \frac{\partial^2 f}{\partial y^2} = 8\ , \quad \frac{\partial^2 f}{\partial y\,\partial x} = -4$$

So

$$D = \frac{\partial^2 f}{\partial x^2}(2,1)\frac{\partial^2 f}{\partial y^2}(2,1) - \left(\frac{\partial^2 f}{\partial y\,\partial x}(2,1)\right)^2 = (2)(8)-(-4)^2 = 0$$

and so the test fails. What can be done in this situation? Sometimes it is possible to examine the function to see directly the nature of a critical point. In our case, we see that $f(x,y) \ge 0$ for all (x,y), since $f(x,y)$ is the sum of fourth and second powers of numbers and hence must be nonnegative. But we also see that $f(2,1) = 0$. Thus $f(x,y) \ge 0 = f(2,1)$ for all (x,y), and hence $(2,1)$ is in fact a *global* minimum for f.

Example 2.22. Find all local maxima and minima of $f(x,y) = (x^2+y^2)e^{-(x^2+y^2)}$.

Solution: First find the critical points, i.e. where $\nabla f = \mathbf{0}$. Since

$$\frac{\partial f}{\partial x} = 2x(1-(x^2+y^2))e^{-(x^2+y^2)}$$

$$\frac{\partial f}{\partial y} = 2y(1-(x^2+y^2))e^{-(x^2+y^2)}$$

then the critical points are $(0,0)$ and all points (x,y) on the unit circle $x^2+y^2=1$.

To use Theorem 2.6, we need the second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = 2[1-(x^2+y^2)-2x^2-2x^2(1-(x^2+y^2))]e^{-(x^2+y^2)}$$

$$\frac{\partial^2 f}{\partial y^2} = 2[1-(x^2+y^2)-2y^2-2y^2(1-(x^2+y^2))]e^{-(x^2+y^2)}$$

$$\frac{\partial^2 f}{\partial y\,\partial x} = -4xy[2-(x^2+y^2)]e^{-(x^2+y^2)}$$

At $(0,0)$, we have $D = 4 > 0$ and $\frac{\partial^2 f}{\partial x^2}(0,0) = 2 > 0$, so $(0,0)$ is a local minimum. However, for points (x,y) on the unit circle $x^2+y^2=1$, we have

$$D = (-4x^2e^{-1})(-4y^2e^{-1}) - (-4xye^{-1})^2 = 0$$

and so the test fails. If we look at the graph of $f(x,y)$, as shown in Figure 2.5.2, it looks like we might have a local maximum for (x,y) on the unit circle $x^2+y^2=1$. If we switch to using polar coordinates (r,θ) instead of (x,y) in $\mathbb{R}^2$, where $r^2 = x^2+y^2$, then we see that we can write $f(x,y)$ as a function $g(r)$ of the variable r alone: $g(r) = r^2e^{-r^2}$. Then $g'(r) = 2r(1-r^2)e^{-r^2}$, so it has a critical point at $r=1$, and we can check that $g''(1) = -4e^{-1} < 0$, so the Second Derivative Test from single-variable calculus says that $r=1$ is a local maximum. But $r=1$ corresponds to the unit circle $x^2+y^2=1$. Thus, the points (x,y) on the unit circle $x^2+y^2=1$ are local maximum points for f.

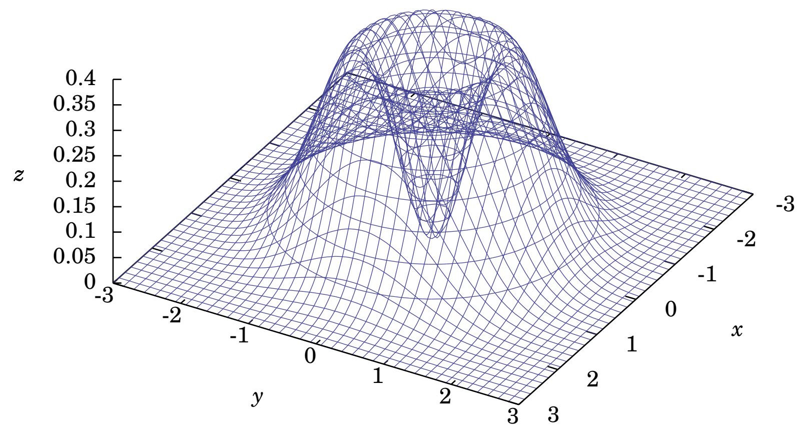


Figure 2.5.2 $f(x,y) = (x^2+y^2)e^{-(x^2+y^2)}$

Exercises

A

For Exercises 1-10, find all local maxima and minima of the function $f(x,y)$.

1. $f(x,y) = x^3 - 3x + y^2$

2. $f(x,y) = x^3 - 12x + y^2 + 8y$

3. $f(x,y) = x^3 - 3x + y^3 - 3y$

4. $f(x,y) = x^3 + 3x^2 + y^3 - 3y^2$

5. $f(x,y) = 2x^3 + 6xy + 3y^2$

6. $f(x,y) = 2x^3 - 6xy + y^2$

7. $f(x,y) = \sqrt{x^2+y^2}$

8. $f(x,y) = x + 2y$

9. $f(x,y) = 4x^2 - 4xy + 2y^2 + 10x - 6y$

10. $f(x,y) = -4x^2 + 4xy - 2y^2 + 16x - 12y$

B

11. For a rectangular solid of volume 1000 cubic meters, find the dimensions that will minimize the surface area. (*Hint: Use the volume condition to write the surface area as a function of just two variables.*)

12. Prove that if (a,b) is a local maximum or local minimum point for a smooth function $f(x,y)$, then the tangent plane to the surface $z = f(x,y)$ at the point $(a,b,f(a,b))$ is parallel to the xy-plane. (*Hint: Use Theorem 2.5.*)

C

13. Find three positive numbers x, y, z whose sum is 10 such that x^2y^2z is a maximum.

2.6 Unconstrained Optimization: Numerical Methods

The types of problems that we solved in the previous section were examples of *unconstrained optimization* problems. That is, we tried to find local (and perhaps even global) maximum and minimum points of real-valued functions $f(x,y)$, where the points (x,y) could be any points in the domain of f. The method we used required us to find the critical points of f, which meant having to solve the equation $\nabla f = \mathbf{0}$, which in general is a system of two equations in two unknowns (x and y). While this was relatively simple for the examples we did, in general this will not be the case. If the equations involve polynomials in x and y of degree three or higher, or complicated expressions involving trigonometric, exponential, or logarithmic functions, then solving even one such equation, let alone two, could be impossible by elementary means.[7]

For example, if one of the equations that had to be solved was

$$x^3 + 9x - 2 = 0 \ ,$$

you may have a hard time getting the exact solutions. Trial and error would not help much, especially since the only real solution[8] turns out to be $\sqrt[3]{\sqrt{28}+1} - \sqrt[3]{\sqrt{28}-1}$. In a situation such as this, the only choice may be to find a solution using some numerical method which gives a sequence of numbers which converge to the actual solution. For example, Newton's method for solving equations $f(x) = 0$, which you probably learned in single-variable calculus. In this section we will describe another method of Newton for finding critical points of real-valued functions of two variables.

Let $f(x,y)$ be a smooth real-valued function, and define

$$D(x,y) \ = \ \frac{\partial^2 f}{\partial x^2}(x,y)\,\frac{\partial^2 f}{\partial y^2}(x,y) - \left(\frac{\partial^2 f}{\partial y\,\partial x}(x,y)\right)^2 \ .$$

Newton's algorithm: Pick an initial point (x_0, y_0). For $n = 0,1,2,3,\ldots$, define:

$$x_{n+1} \ = \ x_n - \frac{\begin{vmatrix} \frac{\partial^2 f}{\partial y^2}(x_n,y_n) & \frac{\partial^2 f}{\partial x\,\partial y}(x_n,y_n) \\ \frac{\partial f}{\partial y}(x_n,y_n) & \frac{\partial f}{\partial x}(x_n,y_n) \end{vmatrix}}{D(x_n,y_n)}, \quad y_{n+1} \ = \ y_n - \frac{\begin{vmatrix} \frac{\partial^2 f}{\partial x^2}(x_n,y_n) & \frac{\partial^2 f}{\partial x\,\partial y}(x_n,y_n) \\ \frac{\partial f}{\partial x}(x_n,y_n) & \frac{\partial f}{\partial y}(x_n,y_n) \end{vmatrix}}{D(x_n,y_n)} \qquad (2.14)$$

Then the sequence of points $(x_n, y_n)_{n=1}^{\infty}$ converges to a critical point. If there are several critical points, then you will have to try different initial points to find them.

[7] This is also a problem for the equivalent method (the Second Derivative Test) in single-variable calculus, though one that is not usually emphasized.

[8] There are also two nonreal, complex number solutions. Cubic polynomial equations in one variable can be solved using *Cardan's formulas*, which are not quite as simple as the familiar quadratic formula. See USPENSKY for more details. There are formulas for solving polynomial equations of degree 4, but it can be proved that there is no general formula for solving equations for polynomials of degree five or higher.

Example 2.23. Find all local maxima and minima of $f(x,y) = x^3 - xy - x + xy^3 - y^4$.
Solution: First calculate the necessary partial derivatives:

$$\frac{\partial f}{\partial x} = 3x^2 - y - 1 + y^3 \ , \qquad \frac{\partial f}{\partial y} = -x + 3xy^2 - 4y^3$$

$$\frac{\partial^2 f}{\partial x^2} = 6x \ , \qquad \frac{\partial^2 f}{\partial y^2} = 6xy - 12y^2 \ , \qquad \frac{\partial^2 f}{\partial y \, \partial x} = -1 + 3y^2$$

Notice that solving $\nabla f = \mathbf{0}$ would involve solving two third-degree polynomial equations in x and y, which in this case can not be done easily.

We need to pick an initial point (x_0, y_0) for our algorithm. Looking at the graph of $z = f(x,y)$ over a large region may help (see Figure 2.6.1 below), though it may be hard to tell where the critical points are.

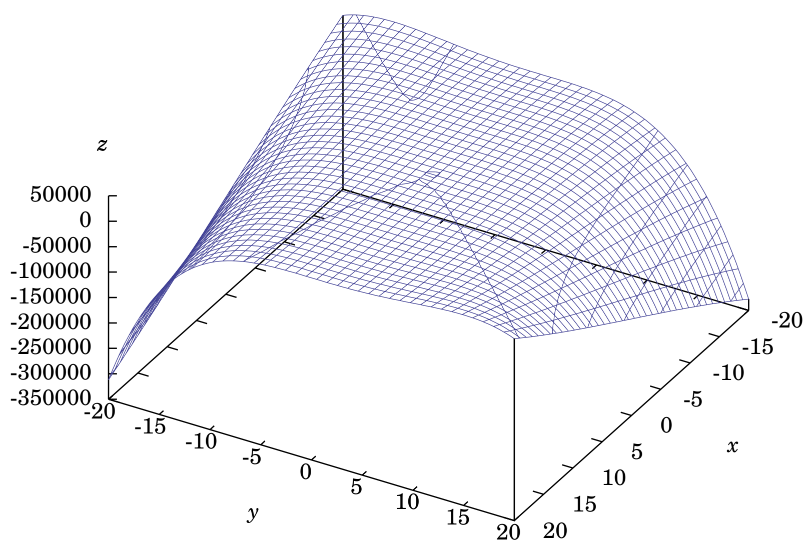


Figure 2.6.1 $f(x,y) = x^3 - xy - x + xy^3 - y^4$ for $-20 \le x \le 20$ and $-20 \le y \le 20$

Notice in the formulas (2.14) that we divide by D, so we should pick an initial point where D is not zero. And we can see that $D(0,0) = (0)(0) - (-1)^2 = -1 \neq 0$, so take $(0,0)$ as our initial point. Since it may take a large number of iterations of Newton's algorithm to be sure that we are close enough to the actual critical point, and since the computations are quite tedious, we will let a computer do the computing. For this, we will write a simple program, using the Java programming language, which will take a given initial point as a parameter and then perform 100 iterations of Newton's algorithm. In each iteration the new point will be printed, so that we can see if there is convergence. The full code is shown in Listing 2.1.

```java
//Program to find the critical points of f(x,y)=x^3-xy-x+xy^3-y^4
public class newton {
 public static void main(String[] args) {
   //Get the initial point (x,y) as command-line parameters
   double x = Double.parseDouble(args[0]); //Initial x value
   double y = Double.parseDouble(args[1]); //Initial y value
   System.out.println("Initial point: (" + x + "," + y + ")");
   //Go through 100 iterations of Newton's algorithm
   for (int n=1; n<=100; n++) {
      double D = fxx(x,y)*fyy(x,y) - Math.pow(fxy(x,y),2);
      double xn = x; double yn = y; //The current x and y values
      if (D == 0) { //We can not divide by 0
         System.out.println("Error: D = 0 at iteration n = " + n);
         System.exit(0); //End the program
      } else { //Calculate the new values for x and y
         x = xn - (fyy(xn,yn)*fx(xn,yn) - fxy(xn,yn)*fy(xn,yn))/D;
         y = yn - (fxx(xn,yn)*fy(xn,yn) - fxy(xn,yn)*fx(xn,yn))/D;
         System.out.println("n = " + n + ": (" + x + "," + y + ")");
      }
   }
 }
 //Below are the parts specific to the function f
 //The first partial derivative of f wrt x: 3x^2-y-1+y^3
 public static double fx(double x, double y) {
   return 3*Math.pow(x,2) - y - 1 + Math.pow(y,3);
 }
 //The first partial derivative of f wrt y: -x+3xy^2-4y^3
 public static double fy(double x, double y) {
   return -x + 3*x*Math.pow(y,2) - 4*Math.pow(y,3);
 }
 //The second partial derivative of f wrt x: 6x
 public static double fxx(double x, double y) {
   return 6*x;
 }
 //The second partial derivative of f wrt y: 6xy-12y^2
 public static double fyy(double x, double y) {
   return 6*x*y - 12*Math.pow(y,2);
 }
 //The mixed second partial derivative of f wrt x and y: -1+3y^2
 public static double fxy(double x, double y) {
   return -1 + 3*Math.pow(y,2);
 }
}
```

Listing 2.1 Program listing for newton.java

To use this program, you should first save the code in Listing 2.1 in a plain text file called `newton.java`. You will need the Java Development Kit[9] to compile the code. In the directory where `newton.java` is saved, run this command at a command prompt to compile the code:
`javac newton.java`
Then run the program with the initial point $(0,0)$ with this command:
`java newton 0 0`

Below is the output of the program using $(0,0)$ as the initial point, truncated to show the first 10 lines and the last 5 lines:

```
java newton 0 0
Initial point: (0.0,0.0)
n = 1: (0.0,-1.0)
n = 2: (1.0,-0.5)
n = 3: (0.6065857885615251,-0.44194107452339687)
n = 4: (0.484506572966545,-0.405341511995805)
n = 5: (0.47123972682634485,-0.3966334583092305)
n = 6: (0.47113558510349535,-0.39636450001936047)
n = 7: (0.4711356343449705,-0.3963643379632247)
n = 8: (0.4711356343449874,-0.39636433796318005)
n = 9: (0.4711356343449874,-0.39636433796318005)
n = 10: (0.4711356343449874,-0.39636433796318005)
...
n = 96: (0.4711356343449874,-0.39636433796318005)
n = 97: (0.4711356343449874,-0.39636433796318005)
n = 98: (0.4711356343449874,-0.39636433796318005)
n = 99: (0.4711356343449874,-0.39636433796318005)
n = 100: (0.4711356343449874,-0.39636433796318005)
```

As you can see, we appear to have converged fairly quickly (after only 8 iterations) to what appears to be an actual critical point (up to Java's level of precision), namely the point $(0.4711356343449874,-0.39636433796318005)$. It is easy to confirm that $\nabla f = \mathbf{0}$ at this point, either by evaluating $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ at the point ourselves or by modifying our program to also print the values of the partial derivatives at the point. It turns out that both partial derivatives are indeed close enough to zero to be considered zero:

$$\frac{\partial f}{\partial x}(0.4711356343449874,-0.39636433796318005) = 4.85722573273506 \times 10^{-17}$$
$$\frac{\partial f}{\partial y}(0.4711356343449874,-0.39636433796318005) = -8.326672684688674 \times 10^{-17}$$

We also have $D(0.4711356343449874,-0.39636433796318005) = -8.776075636032301 < 0$, so by Theorem 2.6 we know that $(0.4711356343449874,-0.39636433796318005)$ is a saddle point.

[9]Available for free at `http://www.oracle.com/technetwork/java/javase/downloads/`

Since ∇f consists of cubic polynomials, it seems likely that there may be three critical points. The computer program makes experimenting with other initial points easy, and trying different values does indeed lead to different sequences which converge:

```
java newton -1 -1
Initial point: (-1.0,-1.0)
n = 1: (-0.5,-0.5)
n = 2: (-0.49295774647887325,-0.08450704225352113)
n = 3: (-0.1855674752461383,-1.2047647348546167)
n = 4: (-0.4540060574531383,-0.8643989895639324)
n = 5: (-0.3672160534444,-0.5426077421319053)
n = 6: (-0.4794622222856417,-0.24529117721011612)
n = 7: (0.11570743992954591,-2.4319791238981274)
n = 8: (-0.05837851765533317,-1.6536079835854451)
n = 9: (-0.129841298650007,-1.121516233310142)
n = 10: (-1.004453014967208,-0.9206128022529645)
n = 11: (-0.5161209914612475,-0.4176293491131443)
n = 12: (-0.5788664043863884,0.2918236503332734)
n = 13: (-0.6985177124230715,0.49848120123515316)
n = 14: (-0.6733618916578702,0.4345777963475479)
n = 15: (-0.6704392913413444,0.4252025996474051)
n = 16: (-0.6703832679150286,0.4250147307973365)
n = 17: (-0.6703832459238701,0.42501465652421205)
n = 18: (-0.6703832459238667,0.4250146565242004)
n = 19: (-0.6703832459238667,0.42501465652420045)
n = 20: (-0.6703832459238667,0.42501465652420045)
...
n = 98: (-0.6703832459238667,0.42501465652420045)
n = 99: (-0.6703832459238667,0.42501465652420045)
n = 100: (-0.6703832459238667,0.42501465652420045)
```

Again, it is easy to confirm that both $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ vanish at the point $(-0.6703832459238667, 0.42501465652420045)$, which means it is a critical point. And

$$D(-0.6703832459238667, 0.42501465652420045) = 15.3853578526055 > 0$$

$$\frac{\partial^2 f}{\partial x^2}(-0.6703832459238667, 0.42501465652420045) = -4.0222994755432 < 0$$

so we know that $(-0.6703832459238667, 0.42501465652420045)$ is a local maximum. An idea of what the graph of f looks like near that point is shown in Figure 2.6.2, which does suggest a local maximum around that point.

Finally, running the computer program with the initial point $(-5, -5)$ yields the critical point $(-7.540962756992551, -5.595509445899435)$, with $D < 0$ at that point, which makes it a saddle point.

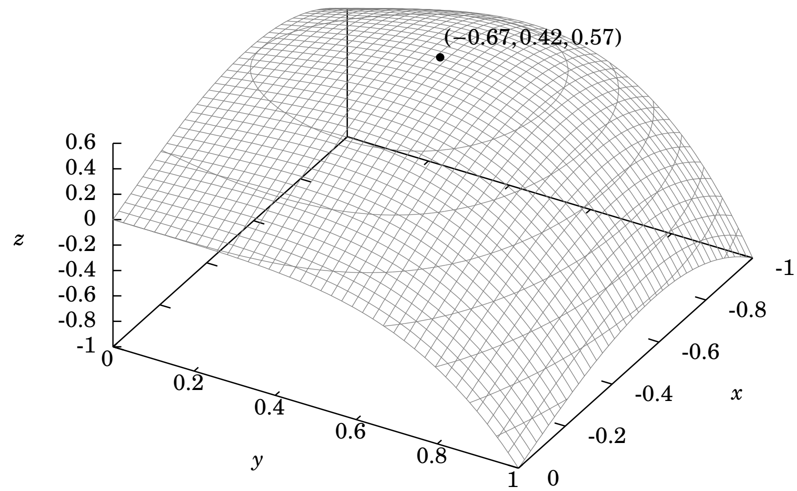


Figure 2.6.2 $f(x,y) = x^3 - xy - x + xy^3 - y^4$ for $-1 \le x \le 0$ and $0 \le y \le 1$

We can summarize our findings for the function $f(x,y) = x^3 - xy - x + xy^3 - y^4$:

$$(0.4711356343449874, -0.39636433796318005): \text{ saddle point}$$
$$(-0.6703832459238667, 0.42501465652420045): \text{ local maximum}$$
$$(-7.540962756992551, -5.595509445899435): \text{ saddle point}$$

The derivation of Newton's algorithm, and the proof that it converges (given a "reasonable" choice for the initial point) requires techniques beyond the scope of this text. See RALSTON and RABINOWITZ for more detail and for discussion of other numerical methods. Our description of Newton's algorithm is the special two-variable case of a more general algorithm that can be applied to functions of $n \ge 2$ variables.

In the case of functions which have a global maximum or minimum, Newton's algorithm can be used to find those points. In general, global maxima and minima tend to be more interesting than local versions, at least in practical applications. A maximization problem can always be turned into a minimization problem (why?), so a large number of methods have been developed to find the global minimum of functions of any number of variables. This field of study is called *nonlinear programming*. Many of these methods are based on the *steepest descent* technique, which is based on an idea that we discussed in Section 2.4. Recall that the negative gradient $-\nabla f$ gives the direction of the fastest rate of decrease of a function f. The crux of the steepest descent idea, then, is that starting from some initial point, you move a certain amount in the direction of $-\nabla f$ at that point. Wherever that takes you

becomes your new point, and you then just keep repeating that procedure until eventually (hopefully) you reach the point where f has its smallest value. There is a "pure" steepest descent method, and a multitude of variations on it that improve the rate of convergence, ease of calculation, etc. In fact, Newton's algorithm can be interpreted as a modified steepest descent method. For more discussion of this, and of nonlinear programming in general, see BAZARAA, SHERALI and SHETTY.

Exercises

C

1. Recall Example 2.21 from the previous section, where we showed that the point $(2,1)$ was a global minimum for the function $f(x,y) = (x-2)^4 + (x-2y)^2$. Notice that our computer program can be modified fairly easily to use this function (just change the return values in the fx, fy, fxx, fyy and fxy function definitions to use the appropriate partial derivative). Either modify that program or write one of your own in a programming language of your choice to show that Newton's algorithm does lead to the point $(2,1)$. First use the initial point $(0,3)$, then use the initial point $(3,2)$, and compare the results. Make sure that your program attempts to do 100 iterations of the algorithm. Did anything strange happen when your program ran? If so, how do you explain it? (*Hint: Something strange should happen.*)

2. There is a version of Newton's algorithm for solving a system of two equations

$$f_1(x,y) = 0 \qquad \text{and} \qquad f_2(x,y) = 0\,,$$

where $f_1(x,y)$ and $f_2(x,y)$ are smooth real-valued functions:
Pick an initial point (x_0,y_0). For $n = 0,1,2,3,\ldots$, define:

$$x_{n+1} \;=\; x_n - \frac{\begin{vmatrix} f_1(x_n,y_n) & f_2(x_n,y_n) \\ \frac{\partial f_1}{\partial y}(x_n,y_n) & \frac{\partial f_2}{\partial y}(x_n,y_n) \end{vmatrix}}{D(x_n,y_n)}\,, \quad y_{n+1} \;=\; y_n + \frac{\begin{vmatrix} f_1(x_n,y_n) & f_2(x_n,y_n) \\ \frac{\partial f_1}{\partial x}(x_n,y_n) & \frac{\partial f_2}{\partial x}(x_n,y_n) \end{vmatrix}}{D(x_n,y_n)}\,, \text{ where}$$

$$D(x_n,y_n) \;=\; \frac{\partial f_1}{\partial x}(x_n,y_n)\,\frac{\partial f_2}{\partial y}(x_n,y_n) - \frac{\partial f_1}{\partial y}(x_n,y_n)\,\frac{\partial f_2}{\partial x}(x_n,y_n)\,.$$

Then the sequence of points $(x_n,y_n)_{n=1}^{\infty}$ converges to a solution. Write a computer program that uses this algorithm to find approximate solutions to the system of equations

$$f_1(x,y) = \sin(xy) - x - y = 0 \qquad \text{and} \qquad f_2(x,y) = e^{2x} - 2x + 3y = 0\,.$$

Show that you get two different solutions when using $(0,0)$ and $(1,1)$ for the initial point (x_0,y_0).

2.7 Constrained Optimization: Lagrange Multipliers

In Sections 2.5 and 2.6 we were concerned with finding maxima and minima of functions without any constraints on the variables (other than being in the domain of the function). What would we do if there were constraints on the variables? The following example illustrates a simple case of this type of problem.

Example 2.24. For a rectangle whose perimeter is 20 m, find the dimensions that will maximize the area.

Solution: The area A of a rectangle with width x and height y is $A = xy$. The perimeter P of the rectangle is then given by the formula $P = 2x+2y$. Since we are given that the perimeter $P = 20$, this problem can be stated as:

$$\begin{aligned}\text{Maximize}: &\; f(x,y) = xy \\ \text{given}: &\; 2x+2y = 20\end{aligned}$$

The reader is probably familiar with a simple method, using single-variable calculus, for solving this problem. Since we must have $2x+2y = 20$, then we can solve for, say, y in terms of x using that equation. This gives $y = 10-x$, which we then substitute into f to get $f(x,y) = xy = x(10-x) = 10x-x^2$. This is now a function of x alone, so we now just have to maximize the function $f(x) = 10x-x^2$ on the interval $[0,10]$. Since $f'(x) = 10-2x = 0 \Rightarrow x = 5$ and $f''(5) = -2 < 0$, then the Second Derivative Test tells us that $x = 5$ is a local maximum for f, and hence $x = 5$ must be the global maximum on the interval $[0,10]$ (since $f = 0$ at the endpoints of the interval). So since $y = 10-x = 5$, then the maximum area occurs for a rectangle whose width and height both are 5 m.

Notice in the above example that the ease of the solution depended on being able to solve for one variable in terms of the other in the equation $2x+2y = 20$. But what if that were not possible (which is often the case)? In this section we will use a general method, called the *Lagrange multiplier method*[10], for solving *constrained optimization* problems:

$$\begin{aligned}\text{Maximize (or minimize)}: &\; f(x,y) \;\text{ (or } f(x,y,z)) \\ \text{given}: &\; g(x,y) = c \;\text{ (or } g(x,y,z) = c) \text{ for some constant } c\end{aligned}$$

The equation $g(x,y) = c$ is called the *constraint equation*, and we say that x and y are *constrained* by $g(x,y) = c$. Points (x,y) which are maxima or minima of $f(x,y)$ with the condition that they satisfy the constraint equation $g(x,y) = c$ are called *constrained maximum* or *constrained minimum* points, respectively. Similar definitions hold for functions of three variables.

The Lagrange multiplier method for solving such problems can now be stated:

[10]Named after the French mathematician Joseph Louis Lagrange (1736-1813).

Theorem 2.7. Let $f(x,y)$ and $g(x,y)$ be smooth functions, and suppose that c is a scalar constant such that $\nabla g(x,y) \neq \mathbf{0}$ for all (x,y) that satisfy the equation $g(x,y) = c$. Then to solve the constrained optimization problem

$$\begin{aligned} \text{Maximize (or minimize)}: &\ f(x,y) \\ \text{given}: &\ g(x,y) = c\ , \end{aligned}$$

find the points (x,y) that solve the equation $\nabla f(x,y) = \lambda \nabla g(x,y)$ for some constant λ (the number λ is called the *Lagrange multiplier*). If there is a constrained maximum or minimum, then it must be such a point.

A rigorous proof of the above theorem requires use of the Implicit Function Theorem, which is beyond the scope of this text.[11] Note that the theorem only gives a *necessary* condition for a point to be a constrained maximum or minimum. Whether a point (x,y) that satisfies $\nabla f(x,y) = \lambda \nabla g(x,y)$ for some λ actually *is* a constrained maximum or minimum can *sometimes* be determined by the nature of the problem itself. For instance, in Example 2.24 it was clear that there had to be a global maximum.

So how can you tell when a point that satisfies the condition in Theorem 2.7 really is a constrained maximum or minimum? The answer is that it depends on the constraint function $g(x,y)$, together with any implicit constraints. It can be shown[12] that if the constraint equation $g(x,y) = c$ (plus any hidden constraints) describes a *bounded* set B in $\mathbb{R}^2$, then the constrained maximum or minimum of $f(x,y)$ will occur either at a point (x,y) satisfying $\nabla f(x,y) = \lambda \nabla g(x,y)$ or at a "boundary" point of the set B.

In Example 2.24 the constraint equation $2x + 2y = 20$ describes a line in $\mathbb{R}^2$, which by itself is not bounded. However, there are "hidden" constraints, due to the nature of the problem, namely $0 \le x, y \le 10$, which cause that line to be restricted to a *line segment* in $\mathbb{R}^2$ (including the endpoints of that line segment), which *is* bounded.

Example 2.25. For a rectangle whose perimeter is 20 m, use the Lagrange multiplier method to find the dimensions that will maximize the area.

Solution: As we saw in Example 2.24, with x and y representing the width and height, respectively, of the rectangle, this problem can be stated as:

$$\begin{aligned} \text{Maximize}: &\ f(x,y) = xy \\ \text{given}: &\ g(x,y) = 2x + 2y = 20 \end{aligned}$$

Then solving the equation $\nabla f(x,y) = \lambda \nabla g(x,y)$ for some λ means solving the equations

[11] See TAYLOR and MANN, § 6.8 for more detail.

[12] Again, see TAYLOR and MANN.

$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x}$ and $\frac{\partial f}{\partial y} = \lambda \frac{\partial g}{\partial y}$, namely:

$$\begin{aligned} y &= 2\lambda\,, \\ x &= 2\lambda \end{aligned}$$

The general idea is to solve for λ in both equations, then set those expressions equal (since they both equal λ) to solve for x and y. Doing this we get

$$\frac{y}{2} = \lambda = \frac{x}{2} \quad \Rightarrow \quad x = y\,,$$

so now substitute either of the expressions for x or y into the constraint equation to solve for x and y:

$$20 = g(x,y) = 2x+2y = 2x+2x = 4x \quad \Rightarrow \quad x=5 \quad \Rightarrow \quad y=5$$

There must be a maximum area, since the minimum area is 0 and $f(5,5) = 25 > 0$, so the point $(5,5)$ that we found (called a *constrained critical point*) must be the constrained maximum.

$\therefore$ The maximum area occurs for a rectangle whose width and height both are 5 m.

Example 2.26. Find the points on the circle $x^2 + y^2 = 80$ which are closest to and farthest from the point $(1,2)$.

Solution: The distance d from any point (x,y) to the point $(1,2)$ is

$$d = \sqrt{(x-1)^2 + (y-2)^2}\,,$$

and minimizing the distance is equivalent to minimizing the square of the distance. Thus the problem can be stated as:

$$\begin{aligned} \text{Maximize (and minimize)}: \ & f(x,y) = (x-1)^2 + (y-2)^2 \\ \text{given}: \ & g(x,y) = x^2 + y^2 = 80 \end{aligned}$$

Solving $\nabla f(x,y) = \lambda \nabla g(x,y)$ means solving the following equations:

$$\begin{aligned} 2(x-1) &= 2\lambda x\,, \\ 2(y-2) &= 2\lambda y \end{aligned}$$

Note that $x \neq 0$ since otherwise we would get $-2 = 0$ in the first equation. Similarly, $y \neq 0$. So we can solve both equations for λ as follows:

$$\frac{x-1}{x} = \lambda = \frac{y-2}{y} \quad \Rightarrow \quad xy - y = xy - 2x \quad \Rightarrow \quad y = 2x$$

Substituting this into $g(x,y) = x^2 + y^2 = 80$ yields $5x^2 = 80$, so $x = \pm 4$. So the two constrained critical points are $(4,8)$ and $(-4,-8)$. Since $f(4,8) = 45$ and $f(-4,-8) = 125$, and since there must be points on the circle closest to and farthest from $(1,2)$, then it must be the case that $(4,8)$ is the point on the circle closest to $(1,2)$ and $(-4,-8)$ is the farthest from $(1,2)$ (see Figure 2.7.1).

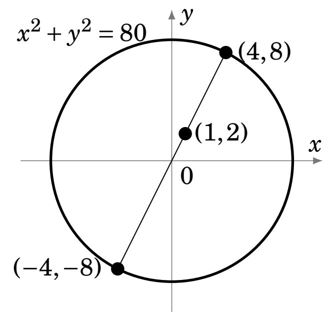


Figure 2.7.1

Notice that since the constraint equation $x^2+y^2 = 80$ describes a circle, which is a bounded set in $\mathbb{R}^2$, then we were guaranteed that the constrained critical points we found were indeed the constrained maximum and minimum.

The Lagrange multiplier method can be extended to functions of three variables.

Example 2.27.

$$\begin{aligned}\text{Maximize (and minimize)}: \; & f(x,y,z) = x+z \\ \text{given}: \; & g(x,y,z) = x^2+y^2+z^2 = 1\end{aligned}$$

Solution: Solve the equation $\nabla f(x,y,z) = \lambda \nabla g(x,y,z)$:

$$\begin{aligned}1 &= 2\lambda x \\ 0 &= 2\lambda y \\ 1 &= 2\lambda z\end{aligned}$$

The first equation implies $\lambda \neq 0$ (otherwise we would have $1 = 0$), so we can divide by λ in the second equation to get $y = 0$ and we can divide by λ in the first and third equations to get $x = \frac{1}{2\lambda} = z$. Substituting these expressions into the constraint equation $g(x,y,z) = x^2 + y^2 + z^2 = 1$ yields the constrained critical points $\left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)$ and $\left(\frac{-1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}\right)$. Since $f\left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right) > f\left(\frac{-1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}\right)$, and since the constraint equation $x^2 + y^2 + z^2 = 1$ describes a sphere (which is bounded) in $\mathbb{R}^3$, then $\left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)$ is the constrained maximum point and $\left(\frac{-1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}\right)$ is the constrained minimum point.

So far we have not attached any significance to the value of the Lagrange multiplier λ. We needed λ only to find the constrained critical points, but made no use of its value. It turns out that λ gives an approximation of the change in the value of the function $f(x,y)$ that we wish to maximize or minimize, when the constant c in the constraint equation $g(x,y) = c$ is changed by 1.

For example, in Example 2.25 we showed that the constrained optimization problem

$$\begin{aligned} \text{Maximize}: \; & f(x,y) = xy \\ \text{given}: \; & g(x,y) = 2x + 2y = 20 \end{aligned}$$

had the solution $(x, y) = (5,5)$, and that $\lambda = x/2 = y/2$. Thus, $\lambda = 2.5$. In a similar fashion we could show that the constrained optimization problem

$$\begin{aligned} \text{Maximize}: \; & f(x,y) = xy \\ \text{given}: \; & g(x,y) = 2x + 2y = 21 \end{aligned}$$

has the solution $(x, y) = (5.25, 5.25)$. So we see that the value of $f(x, y)$ at the constrained maximum increased from $f(5,5) = 25$ to $f(5.25,5.25) = 27.5625$, i.e. it increased by 2.5625 when we increased the value of c in the constraint equation $g(x, y) = c$ from $c = 20$ to $c = 21$. Notice that $\lambda = 2.5$ is close to 2.5625, that is,

$$\lambda \;\approx\; \Delta f \;=\; f(\text{new max. pt}) - f(\text{old max. pt}) \;.$$

Finally, note that solving the equation $\nabla f(x, y) = \lambda \nabla g(x, y)$ means having to solve a system of two (possibly nonlinear) equations in three unknowns, which as we have seen before, may not be possible to do. And the 3-variable case can get even more complicated. All of this somewhat restricts the usefulness of Lagrange's method to relatively simple functions. Luckily there are many numerical methods for solving constrained optimization problems, though we will not discuss them here.[13]

Exercises

A

1. Find the constrained maxima and minima of $f(x, y) = 2x + y$ given that $x^2 + y^2 = 4$.

2. Find the constrained maxima and minima of $f(x, y) = xy$ given that $x^2 + 3y^2 = 6$.

3. Find the points on the circle $x^2 + y^2 = 100$ which are closest to and farthest from the point $(2,3)$.

B

4. Find the constrained maxima and minima of $f(x, y, z) = x + y^2 + 2z$ given that $4x^2 + 9y^2 - 36z^2 = 36$.

5. Find the volume of the largest rectangular parallelepiped that can be inscribed in the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \;.$$

---

[13] See BAZARAA, SHERALI and SHETTY.

3 Multiple Integrals

3.1 Double Integrals

In single-variable calculus, differentiation and integration are thought of as inverse operations. For instance, to integrate a function $f(x)$ it is necessary to find the antiderivative of f, that is, another function $F(x)$ whose derivative is $f(x)$. Is there a similar way of defining integration of real-valued functions of two or more variables? The answer is yes, as we will see shortly. Recall also that the definite integral of a nonnegative function $f(x) \ge 0$ represented the area "under" the curve $y = f(x)$. As we will now see, the *double integral* of a nonnegative real-valued function $f(x, y) \ge 0$ represents the *volume* "under" the surface $z = f(x, y)$.

Let $f(x, y)$ be a continuous function such that $f(x, y) \ge 0$ for all (x, y) on the **rectangle** $R = \{(x, y) : a \le x \le b,\ c \le y \le d\}$ in $\mathbb{R}^2$. We will often write this as $R = [a, b] \times [c, d]$. For any number $x*$ in the interval $[a, b]$, slice the surface $z = f(x, y)$ with the plane $x = x*$ parallel to the yz-plane. Then the trace of the surface in that plane is the *curve* $f(x*, y)$, where $x*$ is fixed and only y varies. The area A under that curve (i.e. the area of the region between the curve and the xy-plane) as y varies over the interval $[c, d]$ then depends only on the value of $x*$. So using the variable x instead of $x*$, let $A(x)$ be that area (see Figure 3.1.1).

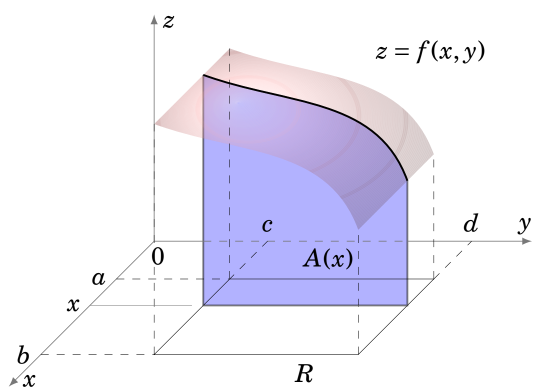


Figure 3.1.1 The area $A(x)$ varies with x

Then $A(x) = \int_c^d f(x, y)\,dy$ since we are treating x as fixed, and only y varies. This makes sense since for a fixed x the function $f(x, y)$ is a continuous function of y over the interval $[c, d]$, so we know that the area under the curve is the definite integral. The area $A(x)$ is a function of x, so by the "slice" or cross-section method from single-variable calculus we know that the volume V of the *solid* under the surface $z = f(x, y)$ but above the xy-plane over the

rectangle R is the integral over $[a,b]$ of that cross-sectional area $A(x)$:

$$V = \int_a^b A(x)\,dx = \int_a^b \left[\int_c^d f(x,y)\,dy \right] dx \tag{3.1}$$

We will always refer to this volume as "the volume under the surface". The above expression uses what are called **iterated integrals**. First the function $f(x,y)$ is integrated as a function of y, treating the variable x as a constant (this is called *integrating with respect to* y). That is what occurs in the "inner" integral between the square brackets in equation (3.1). This is the first iterated integral. Once that integration is performed, the result is then an expression involving only x, which can then be *integrated with respect to* x. That is what occurs in the "outer" integral above (the second iterated integral). The final result is then a number (the volume). This process of going through two iterations of integrals is called *double integration*, and the last expression in equation (3.1) is called a **double integral**.

Notice that integrating $f(x,y)$ with respect to y is the inverse operation of taking the partial derivative of $f(x,y)$ with respect to y. Also, we could just as easily have taken the area of cross-sections under the surface which were parallel to the xz-plane, which would then depend only on the variable y, so that the volume V would be

$$V = \int_c^d \left[\int_a^b f(x,y)\,dx \right] dy\ . \tag{3.2}$$

It turns out that in general[1] the order of the iterated integrals does not matter. Also, we will usually discard the brackets and simply write

$$V = \int_c^d \int_a^b f(x,y)\,dx\,dy\ , \tag{3.3}$$

where it is understood that the fact that dx is written before dy means that the function $f(x,y)$ is first integrated with respect to x using the "inner" limits of integration a and b, and then the resulting function is integrated with respect to y using the "outer" limits of integration c and d. This order of integration can be changed if it is more convenient.

Example 3.1. Find the volume V under the plane $z = 8x + 6y$ over the rectangle $R = [0,1] \times [0,2]$.

[1] due to *Fubini's Theorem*. See Ch. 18 in TAYLOR and MANN.

Solution: We see that $f(x,y) = 8x+6y \ge 0$ for $0 \le x \le 1$ and $0 \le y \le 2$, so:

$$\begin{aligned} V &= \int_0^2 \int_0^1 (8x+6y)\,dx\,dy \\ &= \int_0^2 \left(4x^2 + 6xy \Big|_{x=0}^{x=1} \right) dy \\ &= \int_0^2 (4+6y)\,dy \\ &= 4y + 3y^2 \Big|_0^2 \\ &= 20 \end{aligned}$$

Suppose we had switched the order of integration. We can verify that we still get the same answer:

$$\begin{aligned} V &= \int_0^1 \int_0^2 (8x+6y)\,dy\,dx \\ &= \int_0^1 \left(8xy + 3y^2 \Big|_{y=0}^{y=2} \right) dx \\ &= \int_0^1 (16x+12)\,dx \\ &= 8x^2 + 12x \Big|_0^1 \\ &= 20 \end{aligned}$$

Example 3.2. Find the volume V under the surface $z = e^{x+y}$ over the rectangle $R = [2,3] \times [1,2]$.

Solution: We know that $f(x,y) = e^{x+y} > 0$ for all (x,y), so

$$\begin{aligned} V &= \int_1^2 \int_2^3 e^{x+y}\,dx\,dy \\ &= \int_1^2 \left(e^{x+y} \Big|_{x=2}^{x=3} \right) dy \\ &= \int_1^2 (e^{y+3} - e^{y+2})\,dy \\ &= e^{y+3} - e^{y+2} \Big|_1^2 \\ &= e^5 - e^4 - (e^4 - e^3) \;=\; e^5 - 2e^4 + e^3 \end{aligned}$$

Recall that for a general function $f(x)$, the integral $\int_a^b f(x)\,dx$ represents the difference of the area below the curve $y = f(x)$ but above the x-axis when $f(x) \ge 0$, and the area above the

curve but below the x-axis when $f(x) \le 0$. Similarly, the double integral of any continuous function $f(x,y)$ represents the difference of the volume below the surface $z = f(x,y)$ but above the xy-plane when $f(x,y) \ge 0$, and the volume above the surface but below the xy-plane when $f(x,y) \le 0$. Thus, our method of double integration by means of iterated integrals can be used to evaluate the double integral of *any* continuous function over a rectangle, regardless of whether $f(x,y) \ge 0$ or not.

Example 3.3. Evaluate $\int_0^{2\pi} \int_0^{\pi} \sin(x+y)\,dx\,dy$.

Solution: Note that $f(x,y) = \sin(x+y)$ is both positive and negative over the rectangle $[0,\pi] \times [0,2\pi]$. We can still evaluate the double integral:

$$\begin{aligned}
\int_0^{2\pi} \int_0^{\pi} \sin(x+y)\,dx\,dy &= \int_0^{2\pi} \left(-\cos(x+y)\Big|_{x=0}^{x=\pi} \right) dy \\
&= \int_0^{2\pi} (-\cos(y+\pi) + \cos y)\,dy \\
&= -\sin(y+\pi) + \sin y \Big|_0^{2\pi} \;=\; -\sin 3\pi + \sin 2\pi - (-\sin\pi + \sin 0) \\
&= 0
\end{aligned}$$

Exercises

A

For Exercises 1-4, find the volume under the surface $z = f(x,y)$ over the rectangle R.

1. $f(x,y) = 4xy$, $R = [0,1] \times [0,1]$

2. $f(x,y) = e^{x+y}$, $R = [0,1] \times [-1,1]$

3. $f(x,y) = x^3 + y^2$, $R = [0,1] \times [0,1]$

4. $f(x,y) = x^4 + xy + y^3$, $R = [1,2] \times [0,2]$

For Exercises 5-12, evaluate the given double integral.

5. $\int_0^1 \int_1^2 (1-y)x^2\,dx\,dy$

6. $\int_0^1 \int_0^2 x(x+y)\,dx\,dy$

7. $\int_0^2 \int_0^1 (x+2)\,dx\,dy$

8. $\int_{-1}^2 \int_{-1}^1 x(xy + \sin x)\,dx\,dy$

9. $\int_0^{\pi/2} \int_0^1 xy\cos(x^2 y)\,dx\,dy$

10. $\int_0^{\pi} \int_0^{\pi/2} \sin x \cos(y-\pi)\,dx\,dy$

11. $\int_0^2 \int_1^4 xy\,dx\,dy$

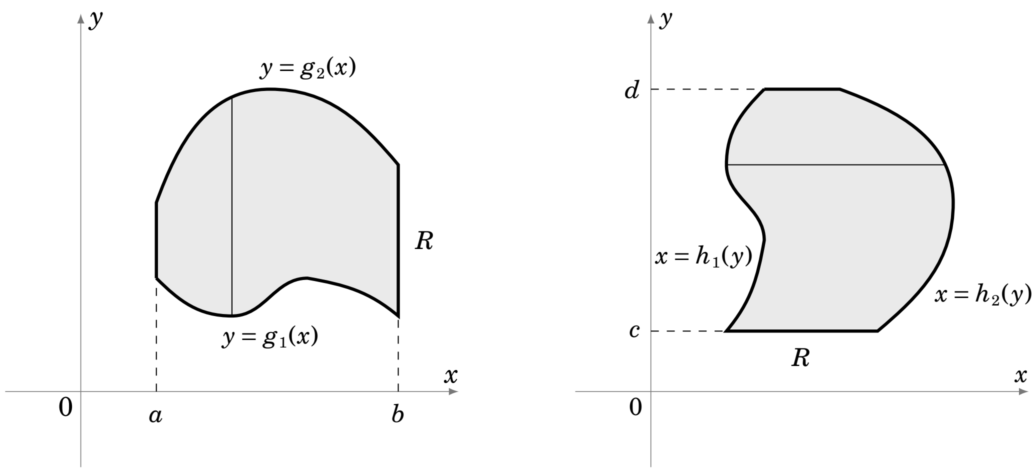
12. $\int_{-1}^1 \int_{-1}^2 1\,dx\,dy$

13. Let M be a constant. Show that $\int_c^d \int_a^b M\,dx\,dy = M(d-c)(b-a)$.

3.2 Double Integrals Over a General Region

In the previous section we got an idea of what a double integral over a rectangle represents. We can now define the double integral of a real-valued function $f(x,y)$ over more general regions in $\mathbb{R}^2$.

Suppose that we have a region R in the xy-plane that is bounded on the left by the vertical line $x = a$, bounded on the right by the vertical line $x = b$ (where $a < b$), bounded below by a curve $y = g_1(x)$, and bounded above by a curve $y = g_2(x)$, as in Figure 3.2.1(a). We will assume that $g_1(x)$ and $g_2(x)$ do not intersect on the open interval (a,b) (they could intersect at the endpoints $x = a$ and $x = b$, though).



(a) Vertical slice: $\int_a^b \int_{g_1(x)}^{g_2(x)} f(x,y)\,dy\,dx$ **(b)** Horizontal slice: $\int_c^d \int_{h_1(y)}^{h_2(y)} f(x,y)\,dx\,dy$

Figure 3.2.1 Double integral over a nonrectangular region R

Then using the slice method from the previous section, the double integral of a real-valued function $f(x,y)$ over the region R, denoted by $\iint\limits_R f(x,y)\,dA$, is given by

$$\iint\limits_R f(x,y)\,dA = \int_a^b \left[\int_{g_1(x)}^{g_2(x)} f(x,y)\,dy \right] dx \tag{3.4}$$

This means that we take vertical slices in the region R between the curves $y = g_1(x)$ and $y = g_2(x)$. The symbol dA is sometimes called an *area element* or *infinitesimal*, with the A signifying area. Note that $f(x,y)$ is first integrated with respect to y, with functions of x as the limits of integration. This makes sense since the result of the first iterated integral will have to be a function of x alone, which then allows us to take the second iterated integral with respect to x.

Similarly, if we have a region R in the xy-plane that is bounded on the left by a curve $x = h_1(y)$, bounded on the right by a curve $x = h_2(y)$, bounded below by the horizontal line

$y = c$, and bounded above by the horizontal line $y = d$ (where $c < d$), as in Figure 3.2.1(b) (assuming that $h_1(y)$ and $h_2(y)$ do not intersect on the open interval (c,d)), then taking horizontal slices gives

$$\iint\limits_R f(x,y)\,dA \;=\; \int_c^d \left[\int_{h_1(y)}^{h_2(y)} f(x,y)\,dx\right] dy \tag{3.5}$$

Notice that these definitions include the case when the region R is a rectangle. Also, if $f(x,y) \ge 0$ for all (x,y) in the region R, then $\iint\limits_R f(x,y)\,dA$ is the volume under the surface $z = f(x,y)$ over the region R.

Example 3.4. Find the volume V under the plane $z = 8x+6y$ over the region $R = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le 2x^2\}$.

Solution: The region R is shown in Figure 3.2.2. Using vertical slices we get:

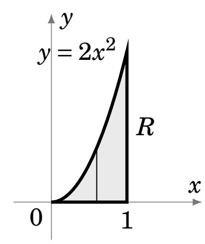


Figure 3.2.2

$$\begin{aligned}
V &= \iint\limits_R (8x+6y)\,dA \\
&= \int_0^1 \left[\int_0^{2x^2} (8x+6y)\,dy\right] dx \\
&= \int_0^1 \left(8xy + 3y^2 \Big|_{y=0}^{y=2x^2}\right) dx \\
&= \int_0^1 (16x^3 + 12x^4)\,dx \\
&= 4x^4 + \tfrac{12}{5}x^5 \Big|_0^1 \;=\; 4 + \tfrac{12}{5} \;=\; \tfrac{32}{5} \;=\; 6.4
\end{aligned}$$

We get the same answer using horizontal slices (see Figure 3.2.3):

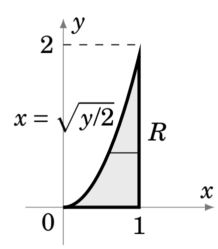


Figure 3.2.3

$$\begin{aligned}
V &= \iint\limits_R (8x+6y)\,dA \\
&= \int_0^2 \left[\int_{\sqrt{y/2}}^1 (8x+6y)\,dx\right] dy \\
&= \int_0^2 \left(4x^2 + 6xy \Big|_{x=\sqrt{y/2}}^{x=1}\right) dy \\
&= \int_0^2 \left(4 + 6y - (2y + \tfrac{6}{\sqrt{2}}y\sqrt{y})\right) dy \;=\; \int_0^2 (4 + 4y - 3\sqrt{2}\,y^{3/2})\,dy \\
&= 4y + 2y^2 - \tfrac{6\sqrt{2}}{5}y^{5/2}\Big|_0^2 \;=\; 8 + 8 - \tfrac{6\sqrt{2}\sqrt{32}}{5} \;=\; 16 - \tfrac{48}{5} \;=\; \tfrac{32}{5} \;=\; 6.4
\end{aligned}$$

Example 3.5. Find the volume V of the solid bounded by the three coordinate planes and the plane $2x + y + 4z = 4$.

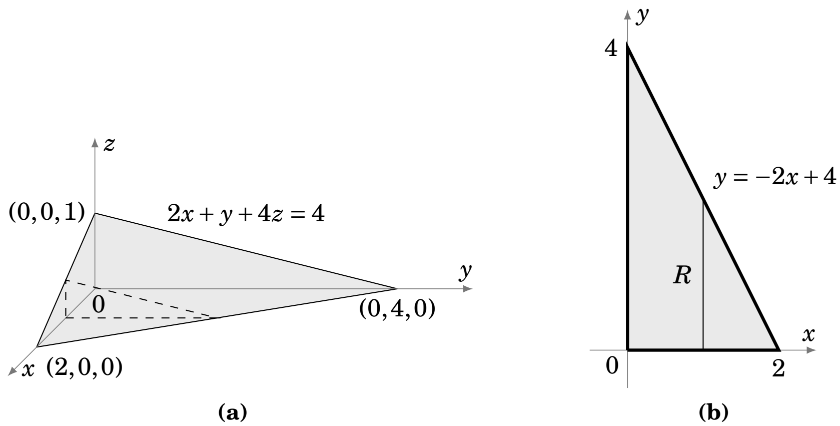


Figure 3.2.4

Solution: The solid is shown in Figure 3.2.4(a) with a typical vertical slice. The volume V is given by $\iint_R f(x,y)\,dA$, where $f(x,y) = z = \frac{1}{4}(4-2x-y)$ and the region R, shown in Figure 3.2.4(b), is $R = \{(x,y) : 0 \le x \le 2,\ 0 \le y \le -2x+4\}$. Using vertical slices in R gives

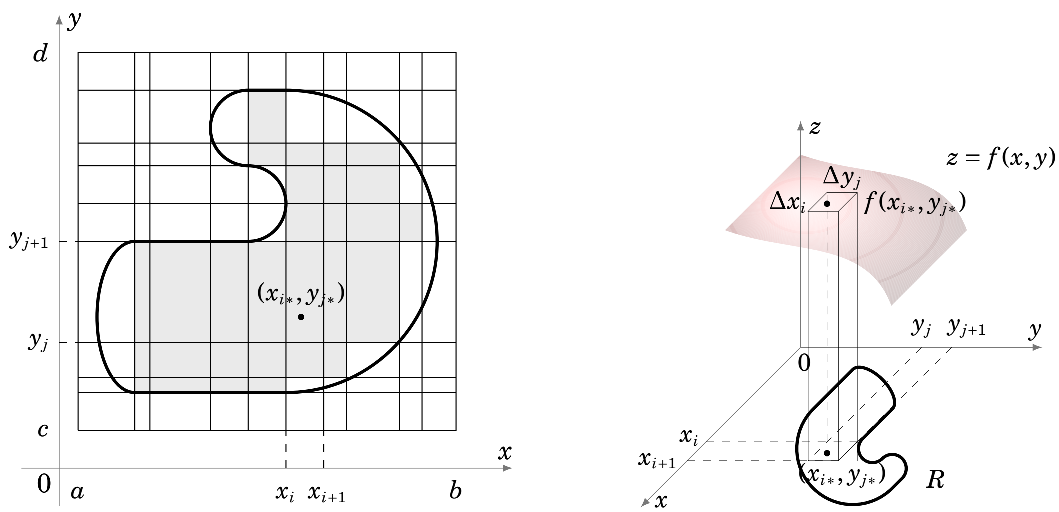
$$
\begin{aligned}
V &= \iint_R \tfrac{1}{4}(4-2x-y)\,dA \\
&= \int_0^2 \left[\int_0^{-2x+4} \tfrac{1}{4}(4-2x-y)\,dy \right] dx \\
&= \int_0^2 \left(-\tfrac{1}{8}(4-2x-y)^2 \Big|_{y=0}^{y=-2x+4} \right) dx \\
&= \int_0^2 \tfrac{1}{8}(4-2x)^2\,dx \\
&= -\tfrac{1}{48}(4-2x)^3 \Big|_0^2 \;=\; \tfrac{64}{48} \;=\; \tfrac{4}{3}
\end{aligned}
$$

For a general region R, which may not be one of the types of regions we have considered so far, the double integral $\iint_R f(x,y)\,dA$ is defined as follows. Assume that $f(x,y)$ is a nonnegative real-valued function and that R is a bounded region in $\mathbb{R}^2$, so it can be enclosed in some rectangle $[a,b]\times[c,d]$. Then divide that rectangle into a grid of subrectangles. Only consider the subrectangles that are enclosed completely within the region R, as shown by the shaded subrectangles in Figure 3.2.5(a). In any such subrectangle $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, pick a point (x_{i*}, y_{j*}). Then the volume under the surface $z = f(x,y)$ over that subrectangle is approximately $f(x_{i*}, y_{j*})\,\Delta x_i\,\Delta y_j$, where $\Delta x_i = x_{i+1} - x_i$, $\Delta y_j = y_{j+1} - y_j$, and $f(x_{i*}, y_{j*})$ is the height and

$\Delta x_i\,\Delta y_j$ is the base area of a parallelepiped, as shown in Figure 3.2.5(b). Then the total volume under the surface is approximately the sum of the volumes of all such parallelepipeds, namely

$$\sum_j \sum_i f(x_{i*}, y_{j*})\,\Delta x_i\,\Delta y_j\ , \tag{3.6}$$

where the summation occurs over the indices of the subrectangles inside R. If we take smaller and smaller subrectangles, so that the length of the largest diagonal of the subrectangles goes to 0, then the subrectangles begin to fill more and more of the region R, and so the above sum approaches the actual volume under the surface $z = f(x, y)$ over the region R. We then *define* $\iint\limits_R f(x, y)\,dA$ as the limit of that double summation (the limit is taken over all subdivisions of the rectangle $[a, b] \times [c, d]$ as the largest diagonal of the subrectangles goes to 0).



(a) Subrectangles inside the region R

(b) Parallelepiped over a subrectangle, with volume $f(x_{i*}, y_{j*})\,\Delta x_i\,\Delta y_j$

Figure 3.2.5 Double integral over a general region R

A similar definition can be made for a function $f(x, y)$ that is not necessarily always non-negative: just replace each mention of volume by the negative volume in the description above when $f(x, y) < 0$. In the case of a region of the type shown in Figure 3.2.1, using the definition of the Riemann integral from single-variable calculus, our definition of $\iint\limits_R f(x, y)\,dA$ reduces to a sequence of two iterated integrals.

Finally, the region R does not have to be bounded. We can evaluate *improper* double integrals (i.e. over an unbounded region, or over a region which contains points where the function $f(x, y)$ is not defined) as a sequence of iterated improper single-variable integrals.

Example 3.6. Evaluate $\int_1^\infty \int_0^{1/x^2} 2y\,dy\,dx$.

Solution:

$$\begin{aligned}\int_1^\infty \int_0^{1/x^2} 2y\,dy\,dx &= \int_1^\infty \left(y^2 \Big|_{y=0}^{y=1/x^2} \right) dx \\ &= \int_1^\infty x^{-4}\,dx = -\tfrac{1}{3}x^{-3}\Big|_1^\infty = 0-(-\tfrac{1}{3}) = \tfrac{1}{3}\end{aligned}$$

Exercises

A

For Exercises 1-6, evaluate the given double integral.

1. $\int_0^1 \int_{\sqrt{x}}^1 24x^2 y\,dy\,dx$

2. $\int_0^\pi \int_0^y \sin x\,dx\,dy$

3. $\int_1^2 \int_0^{\ln x} 4x\,dy\,dx$

4. $\int_0^2 \int_0^{2y} e^{y^2}\,dx\,dy$

5. $\int_0^{\pi/2} \int_0^y \cos x \sin y\,dx\,dy$

6. $\int_0^\infty \int_0^\infty xye^{-(x^2+y^2)}\,dx\,dy$

7. $\int_0^2 \int_0^y 1\,dx\,dy$

8. $\int_0^1 \int_0^{x^2} 2\,dy\,dx$

9. Find the volume V of the solid bounded by the three coordinate planes and the plane $x+y+z=1$.

10. Find the volume V of the solid bounded by the three coordinate planes and the plane $3x+2y+5z=6$.

B

11. Explain why the double integral $\iint\limits_R 1\,dA$ gives the area of the region R. For simplicity, you can assume that R is a region of the type shown in Figure 3.2.1(a).

C

12. Prove that the volume of a tetrahedron with mutually perpendicular adjacent sides of lengths a, b, and c, as in Figure 3.2.6, is $\frac{abc}{6}$. (*Hint: Mimic Example 3.5, and recall from Section 1.5 how three noncollinear points determine a plane.*)

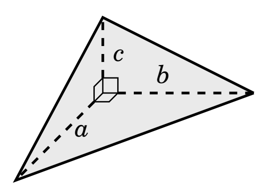


Figure 3.2.6

13. Show how Exercise 12 can be used to solve Exercise 10.

3.3 Triple Integrals

Our definition of a double integral of a real-valued function $f(x,y)$ over a region R in $\mathbb{R}^2$ can be extended to define a *triple integral* of a real-valued function $f(x,y,z)$ over a *solid* S in $\mathbb{R}^3$. We simply proceed as before: the solid S can be enclosed in some rectangular parallelepiped, which is then divided into subparallelepipeds. In each subparallelepiped inside S, with sides of lengths Δx, Δy and Δz, pick a point (x_*, y_*, z_*). Then define the triple integral of $f(x,y,z)$ over S, denoted by $\iiint\limits_S f(x,y,z)\,dV$, by

$$\iiint\limits_S f(x,y,z)\,dV \;=\; \lim \sum\sum\sum f(x_*,y_*,z_*)\,\Delta x\,\Delta y\,\Delta z\ , \tag{3.7}$$

where the limit is over all divisions of the rectangular parallelepiped enclosing S into subparallelepipeds whose largest diagonal is going to 0, and the triple summation is over all the subparallelepipeds inside S. It can be shown that this limit does not depend on the choice of the rectangular parallelepiped enclosing S. The symbol dV is often called the *volume element*.

Physically, what does the triple integral represent? We saw that a double integral could be thought of as the volume under a two-dimensional surface. It turns out that the triple integral simply generalizes this idea: it can be thought of as representing the *hypervolume* under a three-dimensional *hypersurface* $w = f(x,y,z)$ whose graph lies in $\mathbb{R}^4$. In general, the word "volume" is often used as a general term to signify the same concept for any n-dimensional object (e.g. length in $\mathbb{R}^1$, area in $\mathbb{R}^2$). It may be hard to get a grasp on the concept of the "volume" of a four-dimensional object, but at least we now know how to calculate that volume!

In the case where S is a rectangular parallelepiped $[x_1,x_2]\times[y_1,y_2]\times[z_1,z_2]$, that is, $S = \{(x,y,z) : x_1 \le x \le x_2,\ y_1 \le y \le y_2,\ z_1 \le z \le z_2\}$, the triple integral is a sequence of three iterated integrals, namely

$$\iiint\limits_S f(x,y,z)\,dV \;=\; \int_{z_1}^{z_2}\int_{y_1}^{y_2}\int_{x_1}^{x_2} f(x,y,z)\,dx\,dy\,dz\ , \tag{3.8}$$

where the order of integration does not matter. This is the simplest case.

A more complicated case is where S is a solid which is bounded below by a surface $z = g_1(x,y)$, bounded above by a surface $z = g_2(x,y)$, y is bounded between two curves $h_1(x)$ and $h_2(x)$, and x varies between a and b. Then

$$\iiint\limits_S f(x,y,z)\,dV \;=\; \int_a^b \int_{h_1(x)}^{h_2(x)} \int_{g_1(x,y)}^{g_2(x,y)} f(x,y,z)\,dz\,dy\,dx\ . \tag{3.9}$$

Notice in this case that the first iterated integral will result in a function of x and y (since its limits of integration are functions of x and y), which then leaves you with a double integral of

a type that we learned how to evaluate in Section 3.2. There are, of course, many variations on this case (for example, changing the roles of the variables x, y, z), so as you can probably tell, triple integrals can be quite tricky. At this point, just learning how to evaluate a triple integral, regardless of what it represents, is the most important thing. We will see some other ways in which triple integrals are used later in the text.

Example 3.7. Evaluate $\int_0^3 \int_0^2 \int_0^1 (xy+z)\,dx\,dy\,dz$.

Solution:

$$\begin{aligned}
\int_0^3 \int_0^2 \int_0^1 (xy+z)\,dx\,dy\,dz &= \int_0^3 \int_0^2 \left(\tfrac{1}{2}x^2y + xz \Big|_{x=0}^{x=1} \right) dy\,dz \\
&= \int_0^3 \int_0^2 \left(\tfrac{1}{2}y + z \right) dy\,dz \\
&= \int_0^3 \left(\tfrac{1}{4}y^2 + yz \Big|_{y=0}^{y=2} \right) dz \\
&= \int_0^3 (1+2z)\,dz \\
&= z + z^2 \Big|_0^3 \;=\; 12
\end{aligned}$$

Example 3.8. Evaluate $\int_0^1 \int_0^{1-x} \int_0^{2-x-y} (x+y+z)\,dz\,dy\,dx$.

Solution:

$$\begin{aligned}
\int_0^1 \int_0^{1-x} \int_0^{2-x-y} (x+y+z)\,dz\,dy\,dx &= \int_0^1 \int_0^{1-x} \left((x+y)z + \tfrac{1}{2}z^2 \Big|_{z=0}^{z=2-x-y} \right) dy\,dx \\
&= \int_0^1 \int_0^{1-x} \left((x+y)(2-x-y) + \tfrac{1}{2}(2-x-y)^2 \right) dy\,dx \\
&= \int_0^1 \int_0^{1-x} \left(2 - \tfrac{1}{2}x^2 - xy - \tfrac{1}{2}y^2 \right) dy\,dx \\
&= \int_0^1 \left(2y - \tfrac{1}{2}x^2y - xy - \tfrac{1}{2}xy^2 - \tfrac{1}{6}y^3 \Big|_{y=0}^{y=1-x} \right) dx \\
&= \int_0^1 \left(\tfrac{11}{6} - 2x + \tfrac{1}{6}x^3 \right) dx \\
&= \tfrac{11}{6}x - x^2 + \tfrac{1}{24}x^4 \Big|_0^1 \;=\; \tfrac{7}{8}
\end{aligned}$$

Note that the volume V of a solid in $\mathbb{R}^3$ is given by

$$V = \iiint\limits_S 1\,dV\,. \tag{3.10}$$

Since the function being integrated is the constant 1, then the above triple integral reduces to a double integral of the types that we considered in the previous section if the solid is bounded above by some surface $z = f(x,y)$ and bounded below by the xy-plane $z = 0$. There are many other possibilities. For example, the solid could be bounded below and above by surfaces $z = g_1(x,y)$ and $z = g_2(x,y)$, respectively, with y bounded between two curves $h_1(x)$ and $h_2(x)$, and x varies between a and b. Then

$$V = \iiint\limits_S 1\,dV = \int_a^b \int_{h_1(x)}^{h_2(x)} \int_{g_1(x,y)}^{g_2(x,y)} 1\,dz\,dy\,dx = \int_a^b \int_{h_1(x)}^{h_2(x)} (g_2(x,y) - g_1(x,y))\,dy\,dx$$

just like in equation (3.9). See Exercise 10 for an example.

Exercises

A

For Exercises 1-8, evaluate the given triple integral.

1. $\displaystyle\int_0^3 \int_0^2 \int_0^1 xyz\,dx\,dy\,dz$

2. $\displaystyle\int_0^1 \int_0^x \int_0^y xyz\,dz\,dy\,dx$

3. $\displaystyle\int_0^\pi \int_0^x \int_0^{xy} x^2 \sin z\,dz\,dy\,dx$

4. $\displaystyle\int_0^1 \int_0^z \int_0^y ze^{y^2}\,dx\,dy\,dz$

5. $\displaystyle\int_1^e \int_0^y \int_0^{1/y} x^2 z\,dx\,dz\,dy$

6. $\displaystyle\int_1^2 \int_0^{y^2} \int_0^{z^2} yz\,dx\,dz\,dy$

7. $\displaystyle\int_1^2 \int_2^4 \int_0^3 1\,dx\,dy\,dz$

8. $\displaystyle\int_0^1 \int_0^{1-x} \int_0^{1-x-y} 1\,dz\,dy\,dx$

9. Let M be a constant. Show that $\int_{z_1}^{z_2} \int_{y_1}^{y_2} \int_{x_1}^{x_2} M\,dx\,dy\,dz = M(z_2 - z_1)(y_2 - y_1)(x_2 - x_1)$.

B

10. Find the volume V of the solid S bounded by the three coordinate planes, bounded above by the plane $x + y + z = 2$, and bounded below by the plane $z = x + y$.

C

11. Show that $\displaystyle\int_a^b \int_a^z \int_a^y f(x)\,dx\,dy\,dz = \int_a^b \frac{(b-x)^2}{2} f(x)\,dx$. (*Hint: Think of how changing the order of integration in the triple integral changes the limits of integration.*)

3.4 Numerical Approximation of Multiple Integrals

As you have seen, calculating multiple integrals is tricky even for simple functions and regions. For complicated functions, it may not be possible to evaluate one of the iterated integrals in a simple closed form. Luckily there are numerical methods for approximating the value of a multiple integral. The method we will discuss is called the *Monte Carlo method*. The idea behind it is based on the concept of the *average value* of a function, which you learned in single-variable calculus. Recall that for a continuous function $f(x)$, the **average value** $\bar{f}$ of f over an interval $[a,b]$ is defined as

$$\bar{f} = \frac{1}{b-a}\int_a^b f(x)\,dx\,. \tag{3.11}$$

The quantity $b-a$ is the length of the interval $[a,b]$, which can be thought of as the "volume" of the interval. Applying the same reasoning to functions of two or three variables, we define the **average value** of $f(x,y)$ over a region R to be

$$\bar{f} = \frac{1}{A(R)}\iint\limits_R f(x,y)\,dA\,, \tag{3.12}$$

where $A(R)$ is the area of the region R, and we define the **average value** of $f(x,y,z)$ over a solid S to be

$$\bar{f} = \frac{1}{V(S)}\iiint\limits_S f(x,y,z)\,dV\,, \tag{3.13}$$

where $V(S)$ is the volume of the solid S. Thus, for example, we have

$$\iint\limits_R f(x,y)\,dA = A(R)\bar{f}\,. \tag{3.14}$$

The average value of $f(x,y)$ over R can be thought of as representing the sum of all the values of f divided by the number of points in R. Unfortunately there are an infinite number (in fact, *uncountably* many) points in any region, i.e. they can not be listed in a discrete sequence. But what if we took a very large number N of *random* points in the region R (which can be generated by a computer) and then took the average of the values of f for those points, and used that average as the value of $\bar{f}$? This is exactly what the Monte Carlo method does. So in formula (3.14) the approximation we get is

$$\iint\limits_R f(x,y)\,dA \approx A(R)\bar{f} \pm A(R)\sqrt{\frac{\overline{f^2}-(\bar{f})^2}{N}}\,, \tag{3.15}$$

where

$$\bar{f} = \frac{\sum_{i=1}^N f(x_i,y_i)}{N} \quad \text{and} \quad \overline{f^2} = \frac{\sum_{i=1}^N (f(x_i,y_i))^2}{N}\,, \tag{3.16}$$

with the sums taken over the N random points (x_1, y_1), ..., (x_N, y_N). The $\pm$ "error term" in formula (3.15) does not really provide hard bounds on the approximation. It represents a single *standard deviation* from the *expected* value of the integral. That is, it provides a *likely* bound on the error. Due to its use of random points, the Monte Carlo method is an example of a *probabilistic* method (as opposed to *deterministic* methods such as Newton's method, which use a specific formula for generating points).

For example, we can use formula (3.15) to approximate the volume V under the plane $z = 8x + 6y$ over the rectangle $R = [0,1] \times [0,2]$. In Example 3.1 in Section 3.1, we showed that the actual volume is 20. Below is a code listing (montecarlo.java) for a Java program that calculates the volume, using a number of points N that is passed on the command line as a parameter.

```
//Program to approximate the double integral of f(x,y)=8x+6y
//over the rectangle [0,1]x[0,2].
public class montecarlo {
 public static void main(String[] args) {
   //Get the number N of random points as a command-line parameter
   int N = Integer.parseInt(args[0]);
   double x = 0; //x-coordinate of a random point
   double y = 0; //y-coordinate of a random point
   double f = 0.0; //Value of f at a random point
   double mf = 0.0; //Mean of the values of f
   double mf2 = 0.0; //Mean of the values of f^2
   for (int i=0;i<N;i++) { //Get the random coordinates
      x = Math.random(); //x is between 0 and 1
      y = 2 * Math.random(); //y is between 0 and 2
      f = 8*x + 6*y; //Value of the function
      mf = mf + f; //Add to the sum of the f values
      mf2 = mf2 + f*f; //Add to the sum of the f^2 values
   }
   mf = mf/N; //Compute the mean of the f values
   mf2 = mf2/N; //Compute the mean of the f^2 values
   System.out.println("N = " + N + ": integral = " + vol()*mf + " +/- "
    + vol()*Math.sqrt((mf2 - Math.pow(mf,2))/N)); //Print the result
 }
 //The volume of the rectangle [0,1]x[0,2]
 public static double vol() {
   return 1*2;
 }
}
```

Listing 3.1 Program listing for montecarlo.java

The results of running this program with various numbers of random points (e.g. `java montecarlo 100`) are shown below:

```
N = 10:       19.36543087722646  +/- 2.7346060413546147
N = 100:      21.334419561385353 +/- 0.7547037194998519
N = 1000:     19.807662237526227 +/- 0.26701709691370235
N = 10000:    20.080975812043256 +/- 0.08378816229769506
N = 100000:   20.009403854556716 +/- 0.026346782289498317
N = 1000000:  20.000866994982314 +/- 0.008321168748642816
```

As you can see, the approximation is fairly good. As $N \to \infty$, it can be shown that the Monte Carlo approximation converges to the actual volume (on the order of $O(\sqrt{N})$, in computational complexity terminology).

In the above example the region R was a rectangle. To use the Monte Carlo method for a nonrectangular (bounded) region R, only a slight modification is needed. Pick a rectangle $\tilde{R}$ that encloses R, and generate random points in that rectangle as before. Then use those points in the calculation of $\bar{f}$ only if they are inside R. There is no need to calculate the area of R for formula (3.15) in this case, since the exclusion of points not inside R allows you to use the area of the rectangle $\tilde{R}$ instead, similar to before.

For instance, in Example 3.4 we showed that the volume under the surface $z = 8x+6y$ over the nonrectangular region $R = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le 2x^2\}$ is 6.4. Since the rectangle $\tilde{R} = [0,1] \times [0,2]$ contains R, we can use the same program as before, with the only change being a check to see if $y < 2x^2$ for a random point (x,y) in $[0,1] \times [0,2]$. Listing 3.2 below contains the code (montecarlo2.java):

```java
//Program to approximate the double integral of f(x,y)=8x+6y over the
//region bounded by x=0, x=1, y=0, and y=2x^2
public class montecarlo2 {
 public static void main(String[] args) {
   //Get the number N of random points as a command-line parameter
   int N = Integer.parseInt(args[0]);
   double x = 0; //x-coordinate of a random point
   double y = 0; //y-coordinate of a random point
   double f = 0.0; //Value of f at a random point
   double mf = 0.0; //Mean of the values of f
   double mf2 = 0.0; //Mean of the values of f^2
   for (int i=0;i<N;i++) { //Get the random coordinates
      x = Math.random(); //x is between 0 and 1
      y = 2 * Math.random(); //y is between 0 and 2
      if (y < 2*Math.pow(x,2)) { //The point is in the region
         f = 8*x + 6*y; //Value of the function
         mf = mf + f; //Add to the sum of the f values
         mf2 = mf2 + f*f; //Add to the sum of the f^2 values
      }
   }
   mf = mf/N; //Compute the mean of the f values
   mf2 = mf2/N; //Compute the mean of the f^2 values
   System.out.println("N = " + N + ": integral = " + vol()*mf +
```

```
    " +/- " + vol()*Math.sqrt((mf2 - Math.pow(mf,2))/N));
  }
  //The volume of the rectangle [0,1]x[0,2]
  public static double vol() {
    return 1*2;
  }
}
```

Listing 3.2 Program listing for montecarlo2.java

The results of running the program with various numbers of random points (e.g. `java montecarlo2 1000`) are shown below:

```
N = 10:      integral = 6.95747529014894 +/- 2.9185131565120592
N = 100:     integral = 6.3149056229650355 +/- 0.9549009662159909
N = 1000:    integral = 6.477032813858756 +/- 0.31916837260973624
N = 10000:   integral = 6.349975080015089 +/- 0.10040086346895105
N = 100000:  integral = 6.440184132811864 +/- 0.03200476870881392
N = 1000000: integral = 6.417050897922222 +/- 0.01009454409789472
```

To use the Monte Carlo method to evaluate triple integrals, you will need to generate random triples (x, y, z) in a parallelepiped, instead of random pairs (x, y) in a rectangle, and use the volume of the parallelepiped instead of the area of a rectangle in formula (3.15) (see Exercise 2). For a more detailed discussion of numerical integration methods, see PRESS et al.

Exercises

C

1. Write a program that uses the Monte Carlo method to approximate the double integral $\iint\limits_R e^{xy}\,dA$, where $R = [0,1] \times [0,1]$. Show the program output for $N = 10$, 100, 1000, 10000, 100000 and 1000000 random points.

2. Write a program that uses the Monte Carlo method to approximate the triple integral $\iiint\limits_S e^{xyz}\,dV$, where $S = [0,1] \times [0,1] \times [0,1]$. Show the program output for $N = 10$, 100, 1000, 10000, 100000 and 1000000 random points.

3. Repeat Exercise 1 with the region $R = \{(x, y) : -1 \le x \le 1,\ 0 \le y \le x^2\}$.

4. Repeat Exercise 2 with the solid $S = \{(x, y, z) : 0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le z \le 1 - x - y\}$.

5. Use the Monte Carlo method to approximate the volume of a sphere of radius 1.

6. Use the Monte Carlo method to approximate the volume of the ellipsoid $\frac{x^2}{9} + \frac{y^2}{4} + \frac{z^2}{1} = 1$.

3.5 Change of Variables in Multiple Integrals

Given the difficulty of evaluating multiple integrals, the reader may be wondering if it is possible to simplify those integrals using a suitable substitution for the variables. The answer is yes, though it is a bit more complicated than the substitution method which you learned in single-variable calculus.

Recall that if you are given, for example, the definite integral

$$\int_1^2 x^3 \sqrt{x^2-1}\, dx \,,$$

then you would make the substitution

$$\begin{aligned} u &= x^2 - 1 \;\Rightarrow\; x^2 = u+1 \\ du &= 2x\, dx \end{aligned}$$

which changes the limits of integration

$$\begin{aligned} x = 1 &\;\Rightarrow\; u = 0 \\ x = 2 &\;\Rightarrow\; u = 3 \end{aligned}$$

so that we get

$$\begin{aligned} \int_1^2 x^3 \sqrt{x^2-1}\, dx &= \int_1^2 \tfrac{1}{2}x^2 \cdot 2x\sqrt{x^2-1}\, dx \\ &= \int_0^3 \tfrac{1}{2}(u+1)\sqrt{u}\, du \\ &= \tfrac{1}{2}\int_0^3 \left(u^{3/2}+u^{1/2}\right) du \quad \text{, which can be easily integrated to give} \\ &= \tfrac{14\sqrt{3}}{5} \,. \end{aligned}$$

Let us take a different look at what happened when we did that substitution, which will give some motivation for how substitution works in multiple integrals. First, we let $u = x^2 - 1$. On the interval of integration $[1,2]$, the function $x \mapsto x^2 - 1$ is strictly increasing (and maps $[1,2]$ onto $[0,3]$) and hence has an inverse function (defined on the interval $[0,3]$). That is, on $[0,3]$ we can define x as a function of u, namely

$$x \;=\; g(u) \;=\; \sqrt{u+1} \,.$$

Then substituting that expression for x into the function $f(x) = x^3\sqrt{x^2-1}$ gives

$$f(x) \;=\; f(g(u)) \;=\; (u+1)^{3/2}\sqrt{u} \,,$$

and we see that

$$\frac{dx}{du} = g'(u) \Rightarrow dx = g'(u)\,du$$
$$dx = \tfrac{1}{2}(u+1)^{-1/2}\,du\ ,$$

so since

$$g(0) = 1 \Rightarrow 0 = g^{-1}(1)$$
$$g(3) = 2 \Rightarrow 3 = g^{-1}(2)$$

then performing the substitution as we did earlier gives

$$\begin{aligned}
\int_1^2 f(x)\,dx &= \int_1^2 x^3\sqrt{x^2-1}\,dx \\
&= \int_0^3 \tfrac{1}{2}(u+1)\sqrt{u}\,du \quad \text{, which can be written as} \\
&= \int_0^3 (u+1)^{3/2}\sqrt{u}\cdot\tfrac{1}{2}(u+1)^{-1/2}\,du \quad \text{, which means} \\
\int_1^2 f(x)\,dx &= \int_{g^{-1}(1)}^{g^{-1}(2)} f(g(u))\,g'(u)\,du\ .
\end{aligned}$$

In general, if $x = g(u)$ is a one-to-one, differentiable function from an interval $[c,d]$ (which you can think of as being on the "u-axis") onto an interval $[a,b]$ (on the x-axis), which means that $g'(u) \neq 0$ on the interval (c,d), so that $a = g(c)$ and $b = g(d)$, then $c = g^{-1}(a)$ and $d = g^{-1}(b)$, and

$$\int_a^b f(x)\,dx = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(u))\,g'(u)\,du\ . \tag{3.17}$$

This is called the *change of variable* formula for integrals of single-variable functions, and it is what you were implicitly using when doing integration by substitution. This formula turns out to be a special case of a more general formula which can be used to evaluate multiple integrals. We will state the formulas for double and triple integrals involving real-valued functions of two and three variables, respectively. We will assume that all the functions involved are continuously differentiable and that the regions and solids involved all have "reasonable" boundaries. The proof of the following theorem is beyond the scope of the text.[2]

[2]See TAYLOR and MANN, § 15.32 and § 15.62 for all the details.

Theorem 3.1. Change of Variables Formula for Multiple Integrals
Let $x = x(u,v)$ and $y = y(u,v)$ define a one-to-one mapping of a region R' in the uv-plane onto a region R in the xy-plane such that the determinant

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \tag{3.18}$$

is never 0 in R'. Then

$$\iint_R f(x,y)\,dA(x,y) = \iint_{R'} f(x(u,v),y(u,v))\,|J(u,v)|\,dA(u,v)\,. \tag{3.19}$$

We use the notation $dA(x,y)$ and $dA(u,v)$ to denote the area element in the (x,y) and (u,v) coordinates, respectively.

Similarly, if $x = x(u,v,w)$, $y = y(u,v,w)$ and $z = z(u,v,w)$ define a one-to-one mapping of a solid S' in uvw-space onto a solid S in xyz-space such that the determinant

$$J(u,v,w) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{vmatrix} \tag{3.20}$$

is never 0 in S', then

$$\iiint_S f(x,y,z)\,dV(x,y,z) = \iiint_{S'} f(x(u,v,w),y(u,v,w),z(u,v,w))\,|J(u,v,w)|\,dV(u,v,w)\,. \tag{3.21}$$

The determinant $J(u,v)$ in formula (3.18) is called the **Jacobian** of x and y with respect to u and v, and is sometimes written as

$$J(u,v) = \frac{\partial(x,y)}{\partial(u,v)}\,. \tag{3.22}$$

Similarly, the Jacobian $J(u,v,w)$ of three variables is sometimes written as

$$J(u,v,w) = \frac{\partial(x,y,z)}{\partial(u,v,w)}\,. \tag{3.23}$$

Notice that formula (3.19) is saying that $dA(x,y) = |J(u,v)|\,dA(u,v)$, which you can think of as a two-variable version of the relation $dx = g'(u)\,du$ in the single-variable case.

The following example shows how the change of variables formula is used.

Example 3.9. Evaluate $\iint\limits_R e^{\frac{x-y}{x+y}}\,dA$, where $R = \{(x,y) : x \ge 0, y \ge 0, x+y \le 1\}$.

Solution: First, note that evaluating this double integral *without* using substitution is probably impossible, at least in a closed form. By looking at the numerator and denominator of the exponent of e, we will try the substitution $u = x - y$ and $v = x + y$. To use the change of variables formula (3.19), we need to write both x and y in terms of u and v. So solving for x and y gives $x = \frac{1}{2}(u+v)$ and $y = \frac{1}{2}(v-u)$. In Figure 3.5.1 below, we see how the mapping $x = x(u,v) = \frac{1}{2}(u+v)$, $y = y(u,v) = \frac{1}{2}(v-u)$ maps the region R' onto R in a one-to-one manner.

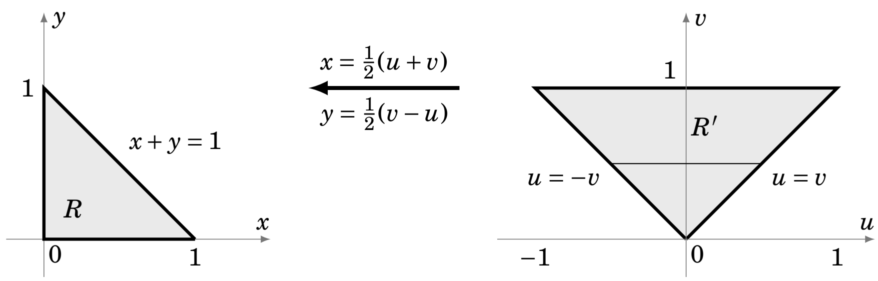


Figure 3.5.1 The regions R and R'

Now we see that

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{vmatrix} = \frac{1}{2} \Rightarrow |J(u,v)| = \left|\frac{1}{2}\right| = \frac{1}{2},$$

so using horizontal slices in R', we have

$$\begin{aligned}
\iint\limits_R e^{\frac{x-y}{x+y}}\,dA &= \iint\limits_{R'} f(x(u,v),y(u,v))\,|J(u,v)|\,dA \\
&= \int_0^1 \int_{-v}^{v} e^{\frac{u}{v}}\,\tfrac{1}{2}\,du\,dv \\
&= \int_0^1 \left(\tfrac{v}{2} e^{\frac{u}{v}} \Big|_{u=-v}^{u=v} \right) dv \\
&= \int_0^1 \tfrac{v}{2}(e - e^{-1})\,dv \\
&= \frac{v^2}{4}(e - e^{-1})\Big|_0^1 = \frac{1}{4}\left(e - \frac{1}{e}\right) = \frac{e^2-1}{4e}
\end{aligned}$$

The change of variables formula can be used to evaluate double integrals in polar coordinates. Letting

$$x = x(r,\theta) = r\cos\theta \quad \text{and} \quad y = y(r,\theta) = r\sin\theta ,$$

we have

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r \Rightarrow |J(u,v)| = |r| = r ,$$

so we have the following formula:

Double Integral in Polar Coordinates

$$\iint\limits_R f(x,y)\,dx\,dy = \iint\limits_{R'} f(r\cos\theta, r\sin\theta)\,r\,dr\,d\theta , \qquad (3.24)$$

where the mapping $x = r\cos\theta$, $y = r\sin\theta$ maps the region R' in the $r\theta$-plane onto the region R in the xy-plane in a one-to-one manner.

Example 3.10. Find the volume V inside the paraboloid $z = x^2 + y^2$ for $0 \le z \le 1$.

Solution: Using vertical slices, we see that

$$V = \iint\limits_R (1-z)\,dA = \iint\limits_R (1-(x^2+y^2))\,dA ,$$

where $R = \{(x,y) : x^2+y^2 \le 1\}$ is the unit disk in $\mathbb{R}^2$ (see Figure 3.5.2). In polar coordinates (r,θ) we know that $x^2+y^2 = r^2$ and that the unit disk R is the set $R' = \{(r,\theta) : 0 \le r \le 1, 0 \le \theta \le 2\pi\}$. Thus,

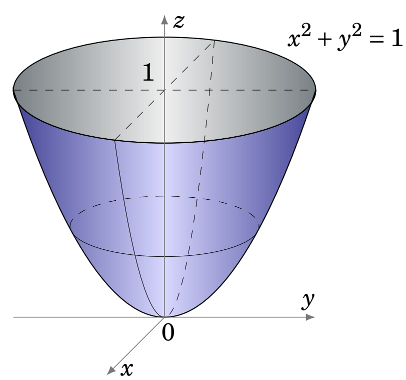


Figure 3.5.2 $z = x^2 + y^2$

$$\begin{aligned} V &= \int_0^{2\pi}\int_0^1 (1-r^2)\,r\,dr\,d\theta \\ &= \int_0^{2\pi}\int_0^1 (r-r^3)\,dr\,d\theta \\ &= \int_0^{2\pi} \left(\frac{r^2}{2} - \frac{r^4}{4} \Big|_{r=0}^{r=1} \right) d\theta \\ &= \int_0^{2\pi} \frac{1}{4}\,d\theta \\ &= \frac{\pi}{2} \end{aligned}$$

Example 3.11. Find the volume V inside the cone $z = \sqrt{x^2+y^2}$ for $0 \le z \le 1$.

Solution: Using vertical slices, we see that

$$V = \iint_R (1-z)\,dA = \iint_R \left(1-\sqrt{x^2+y^2}\right) dA\,,$$

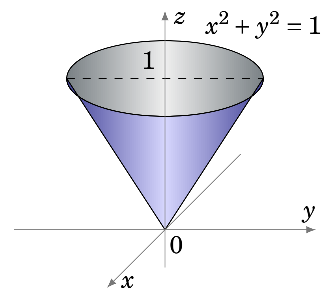


Figure 3.5.3 $z = \sqrt{x^2+y^2}$

where $R = \{(x,y) : x^2+y^2 \le 1\}$ is the unit disk in $\mathbb{R}^2$ (see Figure 3.5.3). In polar coordinates (r,θ) we know that $\sqrt{x^2+y^2} = r$ and that the unit disk R is the set $R' = \{(r,\theta) : 0 \le r \le 1, 0 \le \theta \le 2\pi\}$. Thus,

$$\begin{aligned}
V &= \int_0^{2\pi}\int_0^1 (1-r)\,r\,dr\,d\theta \\
&= \int_0^{2\pi}\int_0^1 (r-r^2)\,dr\,d\theta \\
&= \int_0^{2\pi} \left(\frac{r^2}{2} - \frac{r^3}{3}\Bigg|_{r=0}^{r=1}\right) d\theta \\
&= \int_0^{2\pi} \frac{1}{6}\,d\theta \\
&= \frac{\pi}{3}
\end{aligned}$$

In a similar fashion, it can be shown (see Exercises 5-6) that triple integrals in cylindrical and spherical coordinates take the following forms:

Triple Integral in Cylindrical Coordinates

$$\iiint_S f(x,y,z)\,dx\,dy\,dz = \iiint_{S'} f(r\cos\theta, r\sin\theta, z)\,r\,dr\,d\theta\,dz\,, \tag{3.25}$$

where the mapping $x = r\cos\theta$, $y = r\sin\theta$, $z = z$ maps the solid S' in $r\theta z$-space onto the solid S in xyz-space in a one-to-one manner.

Triple Integral in Spherical Coordinates

$$\iiint_S f(x,y,z)\,dx\,dy\,dz = \iiint_{S'} f(\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi)\,\rho^2\sin\phi\,d\rho\,d\phi\,d\theta\,, \tag{3.26}$$

where the mapping $x = \rho\sin\phi\cos\theta$, $y = \rho\sin\phi\sin\theta$, $z = \rho\cos\phi$ maps the solid S' in $\rho\phi\theta$-space onto the solid S in xyz-space in a one-to-one manner.

Example 3.12. For $a > 0$, find the volume V inside the sphere $S = x^2 + y^2 + z^2 = a^2$.

Solution: We see that S is the set $\rho = a$ in spherical coordinates, so

$$\begin{aligned} V &= \iiint\limits_S 1\,dV = \int_0^{2\pi}\int_0^{\pi}\int_0^{a} 1\,\rho^2 \sin\phi\,d\rho\,d\phi\,d\theta \\ &= \int_0^{2\pi}\int_0^{\pi}\left(\frac{\rho^3}{3}\bigg|_{\rho=0}^{\rho=a}\right)\sin\phi\,d\phi\,d\theta = \int_0^{2\pi}\int_0^{\pi}\frac{a^3}{3}\sin\phi\,d\phi\,d\theta \\ &= \int_0^{2\pi}\left(-\frac{a^3}{3}\cos\phi\bigg|_{\phi=0}^{\phi=\pi}\right)d\theta = \int_0^{2\pi}\frac{2a^3}{3}\,d\theta = \frac{4\pi a^3}{3}\,. \end{aligned}$$

Exercises

A

1. Find the volume V inside the paraboloid $z = x^2 + y^2$ for $0 \le z \le 4$.

2. Find the volume V inside the cone $z = \sqrt{x^2 + y^2}$ for $0 \le z \le 3$.

B

3. Find the volume V of the solid inside both $x^2 + y^2 + z^2 = 4$ and $x^2 + y^2 = 1$.

4. Find the volume V inside both the sphere $x^2 + y^2 + z^2 = 1$ and the cone $z = \sqrt{x^2 + y^2}$.

5. Prove formula (3.25).

6. Prove formula (3.26).

7. Evaluate $\iint\limits_R \sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)dA$, where R is the triangle with vertices $(0,0)$, $(2,0)$ and $(1,1)$. (*Hint: Use the change of variables* $u = (x+y)/2$, $v = (x-y)/2$.)

8. Find the volume of the solid bounded by $z = x^2 + y^2$ and $z^2 = 4(x^2 + y^2)$.

9. Find the volume inside the elliptic cylinder $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ for $0 \le z \le 2$.

C

10. Show that the volume inside the ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ is $\frac{4\pi abc}{3}$. (*Hint: Use the change of variables* $x = au$, $y = bv$, $z = cw$, *then consider Example 3.12.*)

11. Show that the *Beta function*, defined by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt\,, \quad \text{for } x > 0,\ y > 0,$$

satisfies the relation $B(y,x) = B(x,y)$ for $x > 0$, $y > 0$.

12. Using the substitution $t = u/(u+1)$, show that the Beta function can be written as

$$B(x,y) = \int_0^\infty \frac{u^{x-1}}{(u+1)^{x+y}}\,du\,, \quad \text{for } x > 0,\ y > 0.$$

3.6 Application: Center of Mass

Recall from single-variable calculus that for a region $R = \{(x, y) : a \le x \le b, 0 \le y \le f(x)\}$ in $\mathbb{R}^2$ that represents a thin, flat plate (see Figure 3.6.1), where $f(x)$ is a continuous function on $[a, b]$, the *center of mass* of R has coordinates $(\bar{x}, \bar{y})$ given by

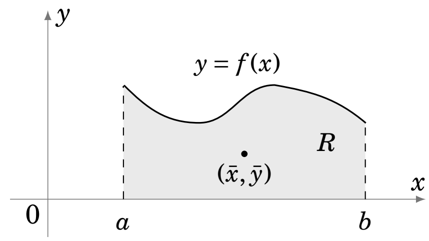


Figure 3.6.1 Center of mass of R

$$\bar{x} = \frac{M_y}{M} \quad \text{and} \quad \bar{y} = \frac{M_x}{M} ,$$

where

$$M_x = \int_a^b \frac{(f(x))^2}{2}\, dx , \quad M_y = \int_a^b x f(x)\, dx , \quad M = \int_a^b f(x)\, dx , \tag{3.27}$$

assuming that R has *uniform density*, i.e the *mass* of R is uniformly distributed over the region. In this case the area M of the region is considered the mass of R (the density is constant, and taken as 1 for simplicity).

In the general case where the density of a region (or *lamina*) R is a continuous function $\delta = \delta(x, y)$ of the coordinates (x, y) of points inside R (where R can be *any* region in $\mathbb{R}^2$) the coordinates $(\bar{x}, \bar{y})$ of the center of mass of R are given by

$$\bar{x} = \frac{M_y}{M} \quad \text{and} \quad \bar{y} = \frac{M_x}{M} , \tag{3.28}$$

where

$$M_y = \iint\limits_R x\delta(x, y)\, dA , \quad M_x = \iint\limits_R y\delta(x, y)\, dA , \quad M = \iint\limits_R \delta(x, y)\, dA , \tag{3.29}$$

The quantities M_x and M_y are called the *moments* (or *first moments*) of the region R about the x-axis and y-axis, respectively. The quantity M is the mass of the region R. To see this, think of taking a small rectangle inside R with dimensions Δx and Δy close to 0. The mass of that rectangle is approximately $\delta(x_*, y_*)\Delta x \Delta y$, for some point (x_*, y_*) in that rectangle. Then the mass of R is the limit of the sums of the masses of all such rectangles inside R as the diagonals of the rectangles approach 0, which is the double integral $\iint\limits_R \delta(x, y)\, dA$.

Note that the formulas in (3.27) represent a special case when $\delta(x, y) = 1$ throughout R in the formulas in (3.29).

Example 3.13. Find the center of mass of the region $R = \{(x, y) : 0 \le x \le 1,\ 0 \le y \le 2x^2\}$, if the density function at (x, y) is $\delta(x, y) = x + y$.

Solution: The region R is shown in Figure 3.6.2. We have

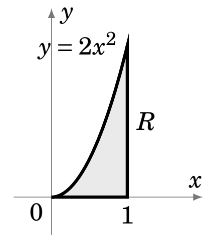


Figure 3.6.2

$$\begin{aligned} M &= \iint_R \delta(x,y)\,dA \\ &= \int_0^1 \int_0^{2x^2} (x+y)\,dy\,dx \\ &= \int_0^1 \left(xy + \frac{y^2}{2} \Bigg|_{y=0}^{y=2x^2} \right) dx \\ &= \int_0^1 (2x^3 + 2x^4)\,dx \\ &= \frac{x^4}{2} + \frac{2x^5}{5} \Bigg|_0^1 = \frac{9}{10} \end{aligned}$$

and

$$\begin{aligned} M_x &= \iint_R y\delta(x,y)\,dA \\ &= \int_0^1 \int_0^{2x^2} y(x+y)\,dy\,dx \\ &= \int_0^1 \left(\frac{xy^2}{2} + \frac{y^3}{3} \Bigg|_{y=0}^{y=2x^2} \right) dx \\ &= \int_0^1 (2x^5 + \frac{8x^6}{3})\,dx \\ &= \frac{x^6}{3} + \frac{8x^7}{21} \Bigg|_0^1 = \frac{5}{7} \end{aligned}$$

$$\begin{aligned} M_y &= \iint_R x\delta(x,y)\,dA \\ &= \int_0^1 \int_0^{2x^2} x(x+y)\,dy\,dx \\ &= \int_0^1 \left(x^2 y + \frac{xy^2}{2} \Bigg|_{y=0}^{y=2x^2} \right) dx \\ &= \int_0^1 (2x^4 + 2x^5)\,dx \\ &= \frac{2x^5}{5} + \frac{x^6}{3} \Bigg|_0^1 = \frac{11}{15}\,, \end{aligned}$$

so the center of mass $(\bar{x}, \bar{y})$ is given by

$$\bar{x} = \frac{M_y}{M} = \frac{11/15}{9/10} = \frac{22}{27}\,, \quad \bar{y} = \frac{M_x}{M} = \frac{5/7}{9/10} = \frac{50}{63}\,.$$

Note how this center of mass is a little further towards the upper corner of the region R than when the density is uniform (it is easy to use the formulas in (3.27) to show that $(\bar{x}, \bar{y}) = \left(\frac{3}{4}, \frac{3}{5}\right)$ in that case). This makes sense since the density function $\delta(x,y) = x + y$ increases as (x,y) approaches that upper corner, where there is quite a bit of area.

In the special case where the density function $\delta(x,y)$ is a constant function on the region R, the center of mass $(\bar{x}, \bar{y})$ is called the *centroid* of R.

The formulas for the center of mass of a region in $\mathbb{R}^2$ can be generalized to a solid S in $\mathbb{R}^3$. Let S be a solid with a continuous mass density function $\delta(x, y, z)$ at any point (x, y, z) in S. Then the center of mass of S has coordinates $(\bar{x}, \bar{y}, \bar{z})$, where

$$\bar{x} = \frac{M_{yz}}{M}\,, \quad \bar{y} = \frac{M_{xz}}{M}\,, \quad \bar{z} = \frac{M_{xy}}{M}\,, \tag{3.30}$$

where

$$M_{yz} = \iiint\limits_S x\delta(x,y,z)\,dV\,, \quad M_{xz} = \iiint\limits_S y\delta(x,y,z)\,dV\,, \quad M_{xy} = \iiint\limits_S z\delta(x,y,z)\,dV\,, \tag{3.31}$$

$$M = \iiint\limits_S \delta(x,y,z)\,dV\,. \tag{3.32}$$

In this case, M_{yz}, M_{xz} and M_{xy} are called the *moments* (or *first moments*) of S around the yz-plane, xz-plane and xy-plane, respectively. Also, M is the mass of S.

Example 3.14. Find the center of mass of the solid $S = \{(x,y,z) : z \ge 0,\ x^2 + y^2 + z^2 \le a^2\}$, if the density function at (x,y,z) is $\delta(x,y,z) = 1$.

Solution: The solid S is just the upper hemisphere inside the sphere of radius a centered at the origin (see Figure 3.6.3). So since the density function is a constant and S is symmetric about the z-axis, then it is clear that $\bar{x} = 0$ and $\bar{y} = 0$, so we need only find $\bar{z}$. We have

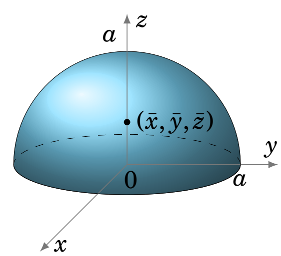


Figure 3.6.3

$$M = \iiint\limits_S \delta(x,y,z)\,dV = \iiint\limits_S 1\,dV = Volume(S).$$

But since the volume of S is half the volume of the sphere of radius a, which we know by Example 3.12 is $\frac{4\pi a^3}{3}$, then $M = \frac{2\pi a^3}{3}$. And

$$\begin{aligned}
M_{xy} &= \iiint\limits_S z\delta(x,y,z)\,dV \\
&= \iiint\limits_S z\,dV \quad \text{, which in spherical coordinates is} \\
&= \int_0^{2\pi}\int_0^{\pi/2}\int_0^a (\rho\cos\phi)\,\rho^2\sin\phi\,d\rho\,d\phi\,d\theta \\
&= \int_0^{2\pi}\int_0^{\pi/2} \sin\phi\cos\phi\left(\int_0^a \rho^3\,d\rho\right)d\phi\,d\theta \\
&= \int_0^{2\pi}\int_0^{\pi/2} \tfrac{a^4}{4}\sin\phi\cos\phi\,d\phi\,d\theta
\end{aligned}$$

$$
\begin{aligned}
M_{xy} &= \int_0^{2\pi} \int_0^{\pi/2} \tfrac{a^4}{8} \sin 2\phi \, d\phi \, d\theta \quad (\text{since } \sin 2\phi = 2 \sin\phi \cos\phi) \\
&= \int_0^{2\pi} \left(-\tfrac{a^4}{16} \cos 2\phi \Big|_{\phi=0}^{\phi=\pi/2} \right) d\theta \\
&= \int_0^{2\pi} \tfrac{a^4}{8} \, d\theta \\
&= \frac{\pi a^4}{4} \, ,
\end{aligned}
$$

so

$$
\bar{z} = \frac{M_{xy}}{M} = \frac{\frac{\pi a^4}{4}}{\frac{2\pi a^3}{3}} = \frac{3a}{8} \, .
$$

Thus, the center of mass of S is $(\bar{x}, \bar{y}, \bar{z}) = \left(0, 0, \frac{3a}{8}\right)$.

Exercises

A

For Exercises 1-5, find the center of mass of the region R with the given density function $\delta(x,y)$.

1. $R = \{(x,y) : 0 \le x \le 2,\ 0 \le y \le 4\}$, $\delta(x,y) = 2y$

2. $R = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le x^2\}$, $\delta(x,y) = x + y$

3. $R = \{(x,y) : y \ge 0,\ x^2 + y^2 \le a^2\}$, $\delta(x,y) = 1$

4. $R = \{(x,y) : y \ge 0,\ x \ge 0,\ 1 \le x^2 + y^2 \le 4\}$, $\delta(x,y) = \sqrt{x^2 + y^2}$

5. $R = \{(x,y) : y \ge 0,\ x^2 + y^2 \le 1\}$, $\delta(x,y) = y$

B

For Exercises 6-10, find the center of mass of the solid S with the given density function $\delta(x,y,z)$.

6. $S = \{(x,y,z) : 0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le z \le 1\}$, $\delta(x,y,z) = xyz$

7. $S = \{(x,y,z) : z \ge 0,\ x^2 + y^2 + z^2 \le a^2\}$, $\delta(x,y,z) = x^2 + y^2 + z^2$

8. $S = \{(x,y,z) : x \ge 0,\ y \ge 0,\ z \ge 0,\ x^2 + y^2 + z^2 \le a^2\}$, $\delta(x,y,z) = 1$

9. $S = \{(x,y,z) : 0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le z \le 1\}$, $\delta(x,y,z) = x^2 + y^2 + z^2$

10. $S = \{(x,y,z) : 0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le z \le 1 - x - y\}$, $\delta(x,y,z) = 1$

3.7 Application: Probability and Expected Value

In this section we will briefly discuss some applications of multiple integrals in the field of probability theory. In particular we will see ways in which multiple integrals can be used to calculate *probabilities* and *expected values*.

Probability

Suppose that you have a standard six-sided (fair) die, and you let a variable X represent the value rolled. Then the *probability* of rolling a 3, written as $P(X = 3)$, is $\frac{1}{6}$, since there are six sides on the die and each one is equally likely to be rolled, and hence in particular the 3 has a one out of six chance of being rolled. Likewise the probability of rolling *at most* a 3, written as $P(X \le 3)$, is $\frac{3}{6} = \frac{1}{2}$, since of the six numbers on the die, there are three equally likely numbers (1, 2, and 3) that are less than or equal to 3. Note that $P(X \le 3) = P(X = 1) + P(X = 2) + P(X = 3)$. We call X a *discrete random variable* on the *sample space* (or *probability space*) Ω consisting of all possible outcomes. In our case, $\Omega = \{1,2,3,4,5,6\}$. An *event* A is a subset of the sample space. For example, in the case of the die, the event $X \le 3$ is the set $\{1,2,3\}$.

Now let X be a variable representing a random real number in the interval $(0,1)$. Note that the set of all real numbers between 0 and 1 is *not* a discrete (or *countable*) set of values, i.e. it can not be put into a one-to-one correspondence with the set of positive integers.[3] In this case, for any real number x in $(0,1)$, it makes no sense to consider $P(X = x)$ since it *must* be 0 (why?). Instead, we consider the probability $P(X \le x)$, which is given by $P(X \le x) = x$. The reasoning is this: the interval $(0,1)$ has length 1, and for x in $(0,1)$ the interval $(0,x)$ has length x. So since X represents a *random* number in $(0,1)$, and hence is *uniformly distributed* over $(0,1)$, then

$$P(X \le x) \;=\; \frac{\text{length of } (0,x)}{\text{length of } (0,1)} \;=\; \frac{x}{1} \;=\; x\;.$$

We call X a *continuous random variable* on the *sample space* $\Omega = (0,1)$. An *event* A is a subset of the sample space. For example, in our case the event $X \le x$ is the set $(0,x)$.

In the case of a discrete random variable, we saw how the probability of an event was the *sum* of the probabilities of the individual outcomes comprising that event (e.g. $P(X \le 3) = P(X = 1) + P(X = 2) + P(X = 3)$ in the die example). For a continuous random variable, the probability of an event will instead be the *integral* of a function, which we will now describe.

Let X be a continuous real-valued random variable on a sample space Ω in $\mathbb{R}$. For sim-

[3]For a proof see p. 9-10 in KAMKE, E., *Theory of Sets*, New York: Dover, 1950.

plicity, let $\Omega = (a, b)$. Define the *distribution function* F of X as

$$F(x) = P(X \le x)\,, \quad \text{for } -\infty < x < \infty \tag{3.33}$$

$$= \begin{cases} 1, & \text{for } x \ge b \\ P(X \le x), & \text{for } a < x < b \\ 0, & \text{for } x \le a\,. \end{cases} \tag{3.34}$$

Suppose that there is a nonnegative, continuous real-valued function f on $\mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(y)\,dy\,, \quad \text{for } -\infty < x < \infty\,, \tag{3.35}$$

and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1\,. \tag{3.36}$$

Then we call f the *probability density function* (or *p.d.f.* for short) for X. We thus have

$$P(X \le x) = \int_{a}^{x} f(y)\,dy\,, \quad \text{for } a < x < b\,. \tag{3.37}$$

Also, by the Fundamental Theorem of Calculus, we have

$$F'(x) = f(x)\,, \quad \text{for } -\infty < x < \infty. \tag{3.38}$$

Example 3.15. Let X represent a randomly selected real number in the interval $(0,1)$. We say that X has the *uniform distribution* on $(0,1)$, with distribution function

$$F(x) = P(X \le x) = \begin{cases} 1, & \text{for } x \ge 1 \\ x, & \text{for } 0 < x < 1 \\ 0, & \text{for } x \le 0\,, \end{cases} \tag{3.39}$$

and probability density function

$$f(x) = F'(x) = \begin{cases} 1, & \text{for } 0 < x < 1 \\ 0, & \text{elsewhere.} \end{cases} \tag{3.40}$$

In general, if X represents a randomly selected real number in an interval (a,b), then X has the uniform distribution function

$$F(x) = P(X \le x) = \begin{cases} 1, & \text{for } x \ge b \\ \frac{x}{b-a}, & \text{for } a < x < b \\ 0, & \text{for } x \le a\,, \end{cases} \tag{3.41}$$

and probability density function

$$f(x) = F'(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a < x < b \\ 0, & \text{elsewhere.} \end{cases} \tag{3.42}$$

Example 3.16. A famous distribution function is given by the *standard normal distribution*, whose probability density function f is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} , \quad \text{for } -\infty < x < \infty. \tag{3.43}$$

This is often called a "bell curve", and is used widely in statistics. Since we are claiming that f is a p.d.f., we *should* have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = 1 \tag{3.44}$$

by formula (3.36), which is equivalent to

$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}\,. \tag{3.45}$$

We can use a double integral in polar coordinates to verify this integral. First,

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy &= \int_{-\infty}^{\infty} e^{-y^2/2}\left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right) dy \\
&= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,dy\right) \\
&= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2
\end{aligned}$$

since the same function is being integrated twice in the middle equation, just with different variables. But using polar coordinates, we see that

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}\,r\,dr\,d\theta \\
&= \int_0^{2\pi}\left(-e^{-r^2/2}\Big|_{r=0}^{r=\infty}\right) d\theta \\
&= \int_0^{2\pi} (0-(-e^0))\,d\theta = \int_0^{2\pi} 1\,d\theta = 2\pi\,,
\end{aligned}$$

and so

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2 = 2\pi \ \text{, and hence}$$

$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}\,.$$

In addition to individual random variables, we can consider *jointly distributed* random variables. For this, we will let X, Y and Z be three real-valued continuous random variables defined on the same sample space Ω in $\mathbb{R}$ (the discussion for two random variables is similar). Then the *joint distribution function* F of X, Y and Z is given by

$$F(x,y,z) = P(X \le x,\ Y \le y,\ Z \le z)\,, \quad \text{for } -\infty < x,y,z < \infty. \tag{3.46}$$

If there is a nonnegative, continuous real-valued function f on $\mathbb{R}^3$ such that

$$F(x,y,z) = \int_{-\infty}^{z}\int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v,w)\,du\,dv\,dw\,, \quad \text{for } -\infty < x,y,z < \infty \tag{3.47}$$

and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z)\,dx\,dy\,dz = 1\,, \tag{3.48}$$

then we call f the *joint probability density function* (or *joint p.d.f.* for short) for X, Y and Z. In general, for $a_1 < b_1$, $a_2 < b_2$, $a_3 < b_3$, we have

$$P(a_1 < X \le b_1,\ a_2 < Y \le b_2,\ a_3 < Z \le b_3) = \int_{a_3}^{b_3}\int_{a_2}^{b_2}\int_{a_1}^{b_1} f(x,y,z)\,dx\,dy\,dz\,, \tag{3.49}$$

with the $\le$ and $<$ symbols interchangeable in any combination. A triple integral, then, can be thought of as representing a probability (for a function f which is a p.d.f.).

Example 3.17. Let a, b, and c be real numbers selected randomly from the interval $(0,1)$. What is the probability that the equation $ax^2 + bx + c = 0$ has at least one real solution x?

Solution: We know by the quadratic formula that there is at least one real solution if $b^2 - 4ac \ge 0$. So we need to calculate $P(b^2 - 4ac \ge 0)$. We will use three jointly distributed random variables to do this. First, since $0 < a,b,c < 1$, we have

$$b^2 - 4ac \ge 0 \Leftrightarrow 0 < 4ac \le b^2 < 1 \Leftrightarrow 0 < 2\sqrt{a}\sqrt{c} \le b < 1\,,$$

where the last relation holds for all $0 < a,c < 1$ such that

$$0 < 4ac < 1 \Leftrightarrow 0 < c < \frac{1}{4a}\,.$$

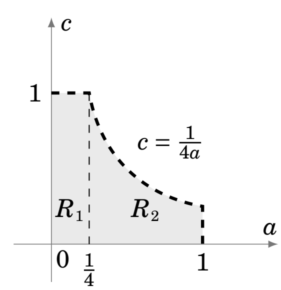


Figure 3.7.1 Region $R = R_1 \cup R_2$

Considering a, b and c as real variables, the region R in the ac-plane where the above relation holds is given by $R = \{(a,c) : 0 < a < 1,\ 0 < c < 1,\ 0 < c < \frac{1}{4a}\}$, which we can see is a union of two regions R_1 and R_2, as in Figure 3.7.1 above.

Now let X, Y and Z be continuous random variables, each representing a randomly selected real number from the interval $(0,1)$ (think of X, Y and Z representing a, b and c,

respectively). Then, similar to how we showed that $f(x) = 1$ is the p.d.f. of the uniform distribution on $(0,1)$, it can be shown that $f(x,y,z) = 1$ for x, y, z in $(0,1)$
(0 elsewhere) is the joint p.d.f. of X, Y and Z. Now,

$$P(b^2 - 4ac \ge 0) \;=\; P((a,c) \in R,\ 2\sqrt{a}\,\sqrt{c} \le b < 1)\,,$$

so this probability is the triple integral of $f(a,b,c) = 1$ as b varies from $2\sqrt{a}\,\sqrt{c}$ to 1 and as (a,c) varies over the region R. Since R can be divided into two regions R_1 and R_2, then the required triple integral can be split into a sum of two triple integrals, using vertical slices in R:

$$\begin{aligned}
P(b^2 - 4ac \ge 0) &= \underbrace{\int_0^{1/4} \int_0^1}_{R_1} \int_{2\sqrt{a}\sqrt{c}}^1 1\,db\,dc\,da \;+\; \underbrace{\int_{1/4}^1 \int_0^{1/4a}}_{R_2} \int_{2\sqrt{a}\sqrt{c}}^1 1\,db\,dc\,da \\
&= \int_0^{1/4} \int_0^1 (1 - 2\sqrt{a}\,\sqrt{c})\,dc\,da \;+\; \int_{1/4}^1 \int_0^{1/4a} (1 - 2\sqrt{a}\,\sqrt{c})\,dc\,da \\
&= \int_0^{1/4} \left(c - \tfrac{4}{3}\sqrt{a}\,c^{3/2} \Big|_{c=0}^{c=1} \right) da \;+\; \int_{1/4}^1 \left(c - \tfrac{4}{3}\sqrt{a}\,c^{3/2} \Big|_{c=0}^{c=1/4a} \right) da \\
&= \int_0^{1/4} \left(1 - \tfrac{4}{3}\sqrt{a}\right) da \;+\; \int_{1/4}^1 \tfrac{1}{12a}\,da \\
&= a - \frac{8}{9}a^{3/2} \Big|_0^{1/4} \;+\; \frac{1}{12}\ln a \Big|_{1/4}^1 \\
&= \left(\frac{1}{4} - \frac{1}{9}\right) + \left(0 - \frac{1}{12}\ln\frac{1}{4}\right) \;=\; \frac{5}{36} + \frac{1}{12}\ln 4 \\
P(b^2 - 4ac \ge 0) &= \frac{5 + 3\ln 4}{36} \;\approx\; 0.2544
\end{aligned}$$

In other words, the equation $ax^2 + bx + c = 0$ has about a 25% chance of being solved!

Expected Value

The *expected value* EX of a random variable X can be thought of as the "average" value of X as it varies over its sample space. If X is a discrete random variable, then

$$EX \;=\; \sum_x x P(X = x)\,, \tag{3.50}$$

with the sum being taken over all elements x of the sample space. For example, if X represents the number rolled on a six-sided die, then

$$EX \;=\; \sum_{x=1}^{6} x P(X = x) \;=\; \sum_{x=1}^{6} x \frac{1}{6} \;=\; 3.5 \tag{3.51}$$

is the expected value of X, which is the average of the integers $1-6$.

If X is a real-valued continuous random variable with p.d.f. f, then

$$EX = \int_{-\infty}^{\infty} x\,f(x)\,dx\,. \tag{3.52}$$

For example, if X has the uniform distribution on the interval $(0,1)$, then its p.d.f. is

$$f(x) = \begin{cases} 1, & \text{for } 0 < x < 1 \\ 0, & \text{elsewhere,} \end{cases} \tag{3.53}$$

and so

$$EX = \int_{-\infty}^{\infty} x\,f(x)\,dx = \int_0^1 x\,dx = \frac{1}{2}\,. \tag{3.54}$$

For a pair of jointly distributed, real-valued continuous random variables X and Y with joint p.d.f. $f(x,y)$, the expected values of X and Y are given by

$$EX = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,f(x,y)\,dx\,dy \quad \text{and} \quad EY = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\,f(x,y)\,dx\,dy\,, \tag{3.55}$$

respectively.

Example 3.18. If you were to pick $n > 2$ random real numbers from the interval $(0,1)$, what are the expected values for the smallest and largest of those numbers?

Solution: Let $U_1,\ldots,U_n$ be n continuous random variables, each representing a randomly selected real number from $(0,1)$, i.e. each has the uniform distribution on $(0,1)$. Define random variables X and Y by

$$X = \min(U_1,\ldots,U_n) \quad \text{and} \quad Y = \max(U_1,\ldots,U_n)\,.$$

Then it can be shown[4] that the joint p.d.f. of X and Y is

$$f(x,y) = \begin{cases} n(n-1)(y-x)^{n-2}, & \text{for } 0 \le x \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases} \tag{3.56}$$

Thus, the expected value of X is

$$\begin{aligned} EX &= \int_0^1 \int_x^1 n(n-1)x(y-x)^{n-2}\,dy\,dx \\ &= \int_0^1 \left(nx(y-x)^{n-1} \Big|_{y=x}^{y=1} \right) dx \\ &= \int_0^1 nx(1-x)^{n-1}\,dx \quad \text{, so integration by parts yields} \\ &= -x(1-x)^n - \frac{1}{n+1}(1-x)^{n+1} \Big|_0^1 \\ EX &= \frac{1}{n+1}\,, \end{aligned}$$

[4]See Ch. 6 in Hoel, Port and Stone.

and similarly (see Exercise 3) it can be shown that

$$EY = \int_0^1 \int_0^y n(n-1)y(y-x)^{n-2}\,dx\,dy = \frac{n}{n+1}.$$

So, for example, if you were to repeatedly take samples of $n = 3$ random real numbers from $(0,1)$, and each time store the minimum and maximum values in the sample, then the average of the minimums would approach $\frac{1}{4}$ and the average of the maximums would approach $\frac{3}{4}$ as the number of samples grows. It would be relatively simple (see Exercise 4) to write a computer program to test this.

Exercises

B

1. Evaluate the integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ using anything you have learned so far.

2. For $\sigma > 0$ and $\mu > 0$, evaluate $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\,dx$.

3. Show that $EY = \frac{n}{n+1}$ in Example 3.18

C

4. Write a computer program (in the language of your choice) that verifies the results in Example 3.18 for the case $n = 3$ by taking large numbers of samples.

5. Repeat Exercise 4 for the case when $n = 4$.

6. For continuous random variables X, Y with joint p.d.f. $f(x, y)$, define the *second moments* $E(X^2)$ and $E(Y^2)$ by

$$E(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f(x,y)\,dx\,dy \quad \text{and} \quad E(Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f(x,y)\,dx\,dy,$$

and the *variances* $\text{Var}(X)$ and $\text{Var}(Y)$ by

$$\text{Var}(X) = E(X^2) - (EX)^2 \quad \text{and} \quad \text{Var}(Y) = E(Y^2) - (EY)^2.$$

Find $\text{Var}(X)$ and $\text{Var}(Y)$ for X and Y as in Example 3.18.

7. Continuing Exercise 6, the *correlation* ρ between X and Y is defined as

$$\rho = \frac{E(XY) - (EX)(EY)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}},$$

where $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy$. Find ρ for X and Y as in Example 3.18. (Note: The quantity $E(XY) - (EX)(EY)$ is called the *covariance* of X and Y.)

8. In Example 3.17 would the answer change if the interval $(0,100)$ is used instead of $(0,1)$? Explain.

4 Line and Surface Integrals

4.1 Line Integrals

In single-variable calculus you learned how to integrate a real-valued function $f(x)$ over an interval $[a,b]$ in $\mathbb{R}^1$. This integral (usually called a *Riemann integral*) can be thought of as an integral over a *path* in $\mathbb{R}^1$, since an interval (or collection of intervals) is really the only kind of "path" in $\mathbb{R}^1$. You may also recall that if $f(x)$ represented the force applied along the x-axis to an object at position x in $[a,b]$, then the *work* W done in moving that object from position $x = a$ to $x = b$ was defined as the integral:

$$W = \int_a^b f(x)\,dx$$

In this section, we will see how to define the integral of a function (either real-valued or vector-valued) of two variables over a general path (i.e. a curve) in $\mathbb{R}^2$. This definition will be motivated by the physical notion of work. We will begin with real-valued functions of two variables.

In physics, the intuitive idea of work is that

$$\text{Work} = \text{Force} \times \text{Distance}\ .$$

Suppose that we want to find the total amount W of work done in moving an object along a curve C in $\mathbb{R}^2$ with a smooth parametrization $x = x(t)$, $y = y(t)$, $a \le t \le b$, with a force $f(x,y)$ which varies with the position (x,y) of the object and is applied in the direction of motion along C (see Figure 4.1.1 below).

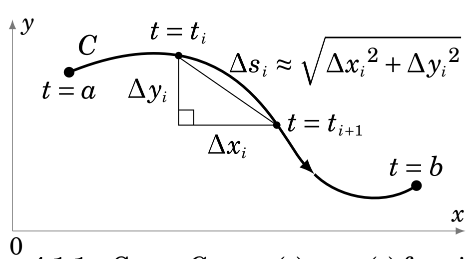


Figure 4.1.1 Curve $C : x = x(t), y = y(t)$ for t in $[a,b]$

We will assume for now that the function $f(x,y)$ is continuous and real-valued, so we only consider the magnitude of the force. Partition the interval $[a,b]$ as follows:

$$a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = b\ , \quad \text{for some integer } n \ge 2$$

As we can see from Figure 4.1.1, over a typical subinterval $[t_i, t_{i+1}]$ the distance Δs_i traveled along the curve is approximately $\sqrt{\Delta x_i^{\,2} + \Delta y_i^{\,2}}$, by the Pythagorean Theorem. Thus, if the subinterval is small enough then the work done in moving the object along that piece of the curve is approximately

$$\text{Force} \times \text{Distance} \;\approx\; f(x_{i*}, y_{i*})\sqrt{\Delta x_i^{\,2} + \Delta y_i^{\,2}}\;, \tag{4.1}$$

where $(x_{i*}, y_{i*}) = (x(t_i*), y(t_i*))$ for some t_i* in $[t_i, t_{i+1}]$, and so

$$W \;\approx\; \sum_{i=0}^{n-1} f(x_{i*}, y_{i*})\sqrt{\Delta x_i^{\,2} + \Delta y_i^{\,2}} \tag{4.2}$$

is approximately the total amount of work done over the entire curve. But since

$$\sqrt{\Delta x_i^{\,2} + \Delta y_i^{\,2}} \;=\; \sqrt{\left(\frac{\Delta x_i}{\Delta t_i}\right)^2 + \left(\frac{\Delta y_i}{\Delta t_i}\right)^2}\;\Delta t_i\;,$$

where $\Delta t_i = t_{i+1} - t_i$, then

$$W \;\approx\; \sum_{i=0}^{n-1} f(x_{i*}, y_{i*})\sqrt{\left(\frac{\Delta x_i}{\Delta t_i}\right)^2 + \left(\frac{\Delta y_i}{\Delta t_i}\right)^2}\;\Delta t_i\;. \tag{4.3}$$

Taking the limit of that sum as the length of the largest subinterval goes to 0, the sum over all subintervals becomes the integral from $t = a$ to $t = b$, $\frac{\Delta x_i}{\Delta t_i}$ and $\frac{\Delta y_i}{\Delta t_i}$ become $x'(t)$ and $y'(t)$, respectively, and $f(x_{i*}, y_{i*})$ becomes $f(x(t), y(t))$, so that

$$W \;=\; \int_a^b f(x(t), y(t))\sqrt{x'(t)^2 + y'(t)^2}\,dt\;. \tag{4.4}$$

The integral on the right side of the above equation gives us our idea of how to define, for *any* real-valued function $f(x, y)$, the integral of $f(x, y)$ along the curve C, called a *line integral*:

Definition 4.1. For a real-valued function $f(x, y)$ and a curve C in $\mathbb{R}^2$, parametrized by $x = x(t)$, $y = y(t)$, $a \le t \le b$, the **line integral of $f(x, y)$ along C with respect to arc length** s is

$$\int_C f(x, y)\,ds \;=\; \int_a^b f(x(t), y(t))\sqrt{x'(t)^2 + y'(t)^2}\,dt\;. \tag{4.5}$$

The symbol ds is the differential of the arc length function

$$s \;=\; s(t) \;=\; \int_a^t \sqrt{x'(u)^2 + y'(u)^2}\,du\;, \tag{4.6}$$

which you may recognize from Section 1.9 as the length of the curve C over the interval $[a,t]$, for all t in $[a,b]$. That is,

$$ds = s'(t)\,dt = \sqrt{x'(t)^2 + y'(t)^2}\,dt\,, \tag{4.7}$$

by the Fundamental Theorem of Calculus.

For a general real-valued function $f(x,y)$, what does the line integral $\int_C f(x,y)\,ds$ represent? The preceding discussion of ds gives us a clue. You can think of differentials as infinitesimal lengths. So if you think of $f(x,y)$ as the height of a picket fence along C, then $f(x,y)\,ds$ can be thought of as approximately the area of a section of that fence over some infinitesimally small section of the curve, and thus the line integral $\int_C f(x,y)\,ds$ is the total area of that picket fence (see Figure 4.1.2).

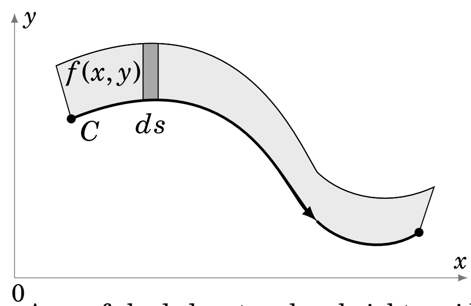


Figure 4.1.2 Area of shaded rectangle = height × width ≈ $f(x,y)\,ds$

Example 4.1. Use a line integral to show that the lateral surface area A of a right circular cylinder of radius r and height h is $2\pi rh$.

Solution: We will use the right circular cylinder with base circle C given by $x^2 + y^2 = r^2$ and with height h in the positive z direction (see Figure 4.1.3). Parametrize C as follows:

$$x = x(t) = r\cos t\,, \quad y = y(t) = r\sin t\,, \quad 0 \le t \le 2\pi$$

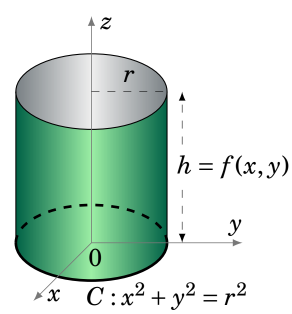


Figure 4.1.3

Let $f(x,y) = h$ for all (x,y). Then

$$\begin{aligned} A &= \int_C f(x,y)\,ds = \int_a^b f(x(t),y(t))\sqrt{x'(t)^2 + y'(t)^2}\,dt \\ &= \int_0^{2\pi} h\sqrt{(-r\sin t)^2 + (r\cos t)^2}\,dt \\ &= h\int_0^{2\pi} r\sqrt{\sin^2 t + \cos^2 t}\,dt \\ &= rh\int_0^{2\pi} 1\,dt = 2\pi rh \end{aligned}$$

Note in Example 4.1 that if we had traversed the circle C twice, i.e. let t vary from 0 to 4π, then we would have gotten an area of $4\pi rh$, i.e. twice the desired area, even though the *curve* itself is still the same (namely, a circle of radius r). Also, notice that we traversed the circle in the counter-clockwise direction. If we had gone in the clockwise direction, using the parametrization

$$x = x(t) = r\cos(2\pi - t)\,, \quad y = y(t) = r\sin(2\pi - t)\,, \quad 0 \le t \le 2\pi\,, \tag{4.8}$$

then it is easy to verify (see Exercise 12) that the value of the line integral is unchanged.

In general, it can be shown (see Exercise 15) that reversing the direction in which a curve C is traversed leaves $\int_C f(x,y)\,ds$ unchanged, for any $f(x,y)$. If a curve C has a parametrization $x = x(t)$, $y = y(t)$, $a \le t \le b$, then denote by $-C$ the same curve as C but traversed in the opposite direction. Then $-C$ is parametrized by

$$x = x(a+b-t)\,, \quad y = y(a+b-t)\,, \quad a \le t \le b\,, \tag{4.9}$$

and we have

$$\int_C f(x,y)\,ds = \int_{-C} f(x,y)\,ds\,. \tag{4.10}$$

Notice that our definition of the line integral was with respect to the arc length parameter s. We can also define

$$\int_C f(x,y)\,dx = \int_a^b f(x(t),y(t))\,x'(t)\,dt \tag{4.11}$$

as the *line integral of $f(x,y)$ along C with respect to x*, and

$$\int_C f(x,y)\,dy = \int_a^b f(x(t),y(t))\,y'(t)\,dt \tag{4.12}$$

as the *line integral of $f(x,y)$ along C with respect to y*.

In the derivation of the formula for a line integral, we used the idea of work as force multiplied by distance. However, we know that force is actually a *vector*. So it would be helpful to develop a vector form for a line integral. For this, suppose that we have a function $\mathbf{f}(x,y)$ defined on $\mathbb{R}^2$ by

$$\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$$

for some continuous real-valued functions $P(x,y)$ and $Q(x,y)$ on $\mathbb{R}^2$. Such a function $\mathbf{f}$ is called a **vector field** on $\mathbb{R}^2$. It is defined at *points* in $\mathbb{R}^2$, and its values are *vectors* in $\mathbb{R}^2$. For a curve C with a smooth parametrization $x = x(t)$, $y = y(t)$, $a \le t \le b$, let

$$\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}$$

be the position vector for a point $(x(t), y(t))$ on C. Then $\mathbf{r}'(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j}$ and so

$$\begin{aligned}
\int_C P(x,y)\,dx + \int_C Q(x,y)\,dy &= \int_a^b P(x(t),y(t))\,x'(t)\,dt + \int_a^b Q(x(t),y(t))\,y'(t)\,dt \\
&= \int_a^b \left(P(x(t),y(t))\,x'(t) + Q(x(t),y(t))\,y'(t)\right)dt \\
&= \int_a^b \mathbf{f}(x(t),y(t)) \cdot \mathbf{r}'(t)\,dt
\end{aligned}$$

by definition of $\mathbf{f}(x, y)$. Notice that the function $\mathbf{f}(x(t), y(t)) \cdot \mathbf{r}'(t)$ is a *real-valued* function on $[a, b]$, so the last integral on the right looks somewhat similar to our earlier definition of a line integral. This leads us to the following definition:

Definition 4.2. For a vector field $\mathbf{f}(x, y) = P(x, y)\mathbf{i} + Q(x, y)\mathbf{j}$ and a curve C with a smooth parametrization $x = x(t)$, $y = y(t)$, $a \le t \le b$, the **line integral of f along** C is

$$\int_C \mathbf{f} \cdot d\mathbf{r} = \int_C P(x,y)\,dx + \int_C Q(x,y)\,dy \tag{4.13}$$

$$= \int_a^b \mathbf{f}(x(t),y(t)) \cdot \mathbf{r}'(t)\,dt\,, \tag{4.14}$$

where $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j}$ is the position vector for points on C.

We use the notation $d\mathbf{r} = \mathbf{r}'(t)\,dt = dx\,\mathbf{i} + dy\,\mathbf{j}$ to denote the **differential** of the vector-valued function $\mathbf{r}$. The line integral in Definition 4.2 is often called a *line integral of a vector field* to distinguish it from the line integral in Definition 4.1 which is called a *line integral of a scalar field*. For convenience we will often write

$$\int_C P(x,y)\,dx + \int_C Q(x,y)\,dy = \int_C P(x,y)\,dx + Q(x,y)\,dy\,,$$

where it is understood that the line integral along C is being applied to both P and Q. The quantity $P(x, y)\,dx + Q(x, y)\,dy$ is known as a **differential form**. For a real-valued function $F(x, y)$, the **differential** of F is $dF = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy$. A differential form $P(x, y)\,dx + Q(x, y)\,dy$ is called **exact** if it equals dF for some function $F(x, y)$.

Recall that if the points on a curve C have position vector $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j}$, then $\mathbf{r}'(t)$ is a tangent vector to C at the point $(x(t), y(t))$ in the direction of increasing t (which we call the *direction of* C). Since C is a smooth curve, then $\mathbf{r}'(t) \ne \mathbf{0}$ on $[a, b]$ and hence

$$\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{\|\mathbf{r}'(t)\|}$$

is the unit tangent vector to C at $(x(t), y(t))$. Putting Definitions 4.1 and 4.2 together we get the following theorem:

Theorem 4.1. For a vector field $\mathbf{f}(x, y) = P(x, y)\,\mathbf{i} + Q(x, y)\,\mathbf{j}$ and a curve C with a smooth parametrization $x = x(t)$, $y = y(t)$, $a \le t \le b$ and position vector $\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}$,

$$\int_C \mathbf{f} \cdot d\mathbf{r} = \int_C \mathbf{f} \cdot \mathbf{T}\, ds \,, \tag{4.15}$$

where $\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{\|\mathbf{r}'(t)\|}$ is the unit tangent vector to C at $(x(t), y(t))$.

If the vector field $\mathbf{f}(x, y)$ represents the force moving an object along a curve C, then the work W done by this force is

$$W = \int_C \mathbf{f} \cdot \mathbf{T}\, ds = \int_C \mathbf{f} \cdot d\mathbf{r} \,. \tag{4.16}$$

Example 4.2. Evaluate $\int_C (x^2 + y^2)\,dx + 2xy\,dy$, where:

(a) $C : x = t \,, \quad y = 2t \,, \quad 0 \le t \le 1$

(b) $C : x = t \,, \quad y = 2t^2 \,, \quad 0 \le t \le 1$

Solution: Figure 4.1.4 shows both curves.

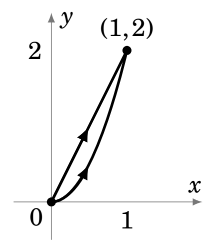


Figure 4.1.4

(a) Since $x'(t) = 1$ and $y'(t) = 2$, then

$$\begin{aligned}
\int_C (x^2 + y^2)\,dx + 2xy\,dy &= \int_0^1 \big((x(t)^2 + y(t)^2)x'(t) + 2x(t)y(t)\,y'(t)\big)\,dt \\
&= \int_0^1 \big((t^2 + 4t^2)(1) + 2t(2t)(2)\big)\,dt \\
&= \int_0^1 13t^2\,dt \\
&= \frac{13t^3}{3}\Bigg|_0^1 = \frac{13}{3}
\end{aligned}$$

(b) Since $x'(t) = 1$ and $y'(t) = 4t$, then

$$\begin{aligned}
\int_C (x^2 + y^2)\,dx + 2xy\,dy &= \int_0^1 \big((x(t)^2 + y(t)^2)x'(t) + 2x(t)y(t)\,y'(t)\big)\,dt \\
&= \int_0^1 \big((t^2 + 4t^4)(1) + 2t(2t^2)(4t)\big)\,dt \\
&= \int_0^1 (t^2 + 20t^4)\,dt \\
&= \frac{t^3}{3} + 4t^5\Bigg|_0^1 = \frac{1}{3} + 4 = \frac{13}{3}
\end{aligned}$$

So in both cases, if the vector field $\mathbf{f}(x,y) = (x^2+y^2)\,\mathbf{i} + 2xy\,\mathbf{j}$ represents the force moving an object from $(0,0)$ to $(1,2)$ along the given curve C, then the work done is $\frac{13}{3}$. This may lead you to think that work (and more generally, the line integral of a vector field) is independent of the path taken. However, as we will see in the next section, this is not always the case.

Although we defined line integrals over a single smooth curve, if C is a *piecewise smooth curve*, that is

$$C = C_1 \cup C_2 \cup \ldots \cup C_n$$

is the union of smooth curves $C_1, \ldots, C_n$, then we can define

$$\int_C \mathbf{f} \cdot d\mathbf{r} = \int_{C_1} \mathbf{f} \cdot d\mathbf{r}_1 + \int_{C_2} \mathbf{f} \cdot d\mathbf{r}_2 + \ldots + \int_{C_n} \mathbf{f} \cdot d\mathbf{r}_n$$

where each $\mathbf{r}_i$ is the position vector of the curve C_i.

Example 4.3. Evaluate $\int_C (x^2+y^2)\,dx + 2xy\,dy$, where C is the polygonal path from $(0,0)$ to $(0,2)$ to $(1,2)$.

Solution: Write $C = C_1 \cup C_2$, where C_1 is the curve given by $x = 0$, $y = t$, $0 \le t \le 2$ and C_2 is the curve given by $x = t$, $y = 2$, $0 \le t \le 1$ (see Figure 4.1.5). Then

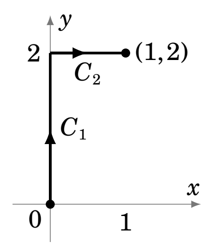


Figure 4.1.5

$$\begin{aligned}
\int_C (x^2+y^2)\,dx + 2xy\,dy &= \int_{C_1} (x^2+y^2)\,dx + 2xy\,dy \\
&\quad + \int_{C_2} (x^2+y^2)\,dx + 2xy\,dy \\
&= \int_0^2 \big((0^2+t^2)(0) + 2(0)t(1)\big)\,dt + \int_0^1 \big((t^2+4)(1) + 2t(2)(0)\big)\,dt \\
&= \int_0^2 0\,dt + \int_0^1 (t^2+4)\,dt \\
&= \frac{t^3}{3} + 4t\,\Bigg|_0^1 = \frac{1}{3} + 4 = \frac{13}{3}
\end{aligned}$$

Line integral notation varies quite a bit. For example, in physics it is common to see the notation $\int_a^b \mathbf{f} \cdot d\mathbf{l}$, where it is understood that the limits of integration a and b are for the underlying parameter t of the curve, and the letter $\mathbf{l}$ signifies length. Also, the formulation $\int_C \mathbf{f} \cdot \mathbf{T}\,ds$ from Theorem 4.1 is often preferred in physics since it emphasizes the idea of integrating the tangential component $\mathbf{f} \cdot \mathbf{T}$ of $\mathbf{f}$ in the direction of $\mathbf{T}$ (i.e. in the direction of C), which is a useful physical interpretation of line integrals.

Exercises

A

For Exercises 1-4, calculate $\int_C f(x,y)\,ds$ for the given function $f(x,y)$ and curve C.

1. $f(x,y) = xy;\quad C: x = \cos t,\ y = \sin t,\ 0 \le t \le \pi/2$

2. $f(x,y) = \dfrac{x}{x^2+1};\quad C: x = t,\ y = 0,\ 0 \le t \le 1$

3. $f(x,y) = 2x + y;\quad C$: polygonal path from $(0,0)$ to $(3,0)$ to $(3,2)$

4. $f(x,y) = x + y^2;\quad C$: path from $(2,0)$ counterclockwise along the circle $x^2 + y^2 = 4$ to the point $(-2,0)$ and then back to $(2,0)$ along the x-axis

5. Use a line integral to find the lateral surface area of the part of the cylinder $x^2 + y^2 = 4$ below the plane $x + 2y + z = 6$ and above the xy-plane.

For Exercises 6-11, calculate $\int_C \mathbf{f}\cdot d\mathbf{r}$ for the given vector field $\mathbf{f}(x,y)$ and curve C.

6. $\mathbf{f}(x,y) = \mathbf{i} - \mathbf{j};\quad C: x = 3t,\ y = 2t,\ 0 \le t \le 1$

7. $\mathbf{f}(x,y) = y\,\mathbf{i} - x\,\mathbf{j};\quad C: x = \cos t,\ y = \sin t,\ 0 \le t \le 2\pi$

8. $\mathbf{f}(x,y) = x\,\mathbf{i} + y\,\mathbf{j};\quad C: x = \cos t,\ y = \sin t,\ 0 \le t \le 2\pi$

9. $\mathbf{f}(x,y) = (x^2 - y)\,\mathbf{i} + (x - y^2)\,\mathbf{j};\quad C: x = \cos t,\ y = \sin t,\ 0 \le t \le 2\pi$

10. $\mathbf{f}(x,y) = xy^2\,\mathbf{i} + xy^3\,\mathbf{j};\quad C$: the polygonal path from $(0,0)$ to $(1,0)$ to $(0,1)$ to $(0,0)$

11. $\mathbf{f}(x,y) = (x^2 + y^2)\,\mathbf{i};\quad C: x = 2 + \cos t,\ y = \sin t,\ 0 \le t \le 2\pi$

B

12. Verify that the value of the line integral in Example 4.1 is unchanged when using the parametrization of the circle C given in formulas (4.8).

13. Show that if $\mathbf{f} \perp \mathbf{r}'(t)$ at each point $\mathbf{r}(t)$ along a smooth curve C, then $\int_C \mathbf{f}\cdot d\mathbf{r} = 0$.

14. Show that if $\mathbf{f}$ points in the same direction as $\mathbf{r}'(t)$ at each point $\mathbf{r}(t)$ along a smooth curve C, then $\int_C \mathbf{f}\cdot d\mathbf{r} = \int_C \|\mathbf{f}\|\,ds$.

C

15. Prove that $\int_C f(x,y)\,ds = \int_{-C} f(x,y)\,ds$. (*Hint: Use formulas (4.9).*)

16. Let C be a smooth curve with arc length L, and suppose that $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ is a vector field such that $\|\mathbf{f}(x,y)\| \le M$ for all (x,y) on C. Show that $\left|\int_C \mathbf{f}\cdot d\mathbf{r}\right| \le ML$. (*Hint: Recall that* $\left|\int_a^b g(x)\,dx\right| \le \int_a^b |g(x)|\,dx$ *for Riemann integrals.*)

17. Prove that the Riemann integral $\int_a^b f(x)\,dx$ is a special case of a line integral.

4.2 Properties of Line Integrals

We know from the previous section that for line integrals of real-valued functions (scalar fields), reversing the direction in which the integral is taken along a curve does not change the value of the line integral:

$$\int_C f(x,y)\,ds = \int_{-C} f(x,y)\,ds \tag{4.17}$$

For line integrals of vector fields, however, the value does change. To see this, let $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ be a vector field, with P and Q continuously differentiable functions. Let C be a smooth curve parametrized by $x = x(t)$, $y = y(t)$, $a \le t \le b$, with position vector $\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}$ (we will usually abbreviate this by saying that $C : \mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}$ is a smooth curve). We know that the curve $-C$ traversed in the opposite direction is parametrized by $x = x(a+b-t)$, $y = y(a+b-t)$, $a \le t \le b$. Then

$$\begin{aligned}
\int_{-C} P(x,y)\,dx &= \int_a^b P(x(a+b-t), y(a+b-t))\,\frac{d}{dt}(x(a+b-t))\,dt \\
&= \int_a^b P(x(a+b-t), y(a+b-t))(-x'(a+b-t))\,dt \quad \text{(by the Chain Rule)} \\
&= \int_b^a P(x(u), y(u))(-x'(u))(-du) \quad \text{(by letting } u = a+b-t\text{)} \\
&= \int_b^a P(x(u), y(u))\,x'(u)\,du \\
&= -\int_a^b P(x(u), y(u))\,x'(u)\,du\ , \quad \text{since } \int_b^a = -\int_a^b\ , \text{ so} \\
\int_{-C} P(x,y)\,dx &= -\int_C P(x,y)\,dx
\end{aligned}$$

since we are just using a different letter (u) for the line integral along C. A similar argument shows that

$$\int_{-C} Q(x,y)\,dy = -\int_C Q(x,y)\,dy\ ,$$

and hence

$$\begin{aligned}
\int_{-C} \mathbf{f}\cdot d\mathbf{r} &= \int_{-C} P(x,y)\,dx + \int_{-C} Q(x,y)\,dy \\
&= -\int_C P(x,y)\,dx + -\int_C Q(x,y)\,dy \\
&= -\left(\int_C P(x,y)\,dx + \int_C Q(x,y)\,dy\right) \\
\int_{-C} \mathbf{f}\cdot d\mathbf{r} &= -\int_C \mathbf{f}\cdot d\mathbf{r}\ .
\end{aligned} \tag{4.18}$$

The above formula can be interpreted in terms of the work done by a force $\mathbf{f}(x,y)$ (treated as a vector) moving an object along a curve C: the total work performed moving the object along C from its initial point to its terminal point, and then back to the initial point moving backwards along the same path, is zero. This is because when force is considered as a vector, direction is accounted for.

The preceding discussion shows the importance of always taking the *direction* of the curve into account when using line integrals of vector fields. For this reason, the curves in line integrals are sometimes referred to as *directed curves* or *oriented curves*.

Recall that our definition of a line integral required that we have *a* parametrization $x = x(t),\ y = y(t),\ a \le t \le b$ for the curve C. But as we know, any curve has infinitely many parametrizations. So could we get a different value for a line integral using some other parametrization of C, say, $x = \tilde{x}(u),\ y = \tilde{y}(u),\ c \le u \le d$? If so, this would mean that our definition is not well-defined. Luckily, it turns out that the value of a line integral of a vector field is unchanged as long as the direction of the curve C is preserved by whatever parametrization is chosen:

Theorem 4.2. Let $\mathbf{f}(x,y) = P(x,y)\mathbf{i} + Q(x,y)\mathbf{j}$ be a vector field, and let C be a smooth curve parametrized by $x = x(t),\ y = y(t),\ a \le t \le b$. Suppose that $t = \alpha(u)$ for $c \le u \le d$, such that $a = \alpha(c)$, $b = \alpha(d)$, and $\alpha'(u) > 0$ on the open interval (c,d) (i.e. $\alpha(u)$ is strictly increasing on $[c,d]$). Then $\int_C \mathbf{f}\cdot d\mathbf{r}$ has the same value for the parametrizations $x = x(t),\ y = y(t),\ a \le t \le b$ and $x = \tilde{x}(u) = x(\alpha(u)),\ y = \tilde{y}(u) = y(\alpha(u)),\ c \le u \le d$.

Proof: Since $\alpha(u)$ is strictly increasing and maps $[c,d]$ onto $[a,b]$, then we know that $t = \alpha(u)$ has an inverse function $u = \alpha^{-1}(t)$ defined on $[a,b]$ such that $c = \alpha^{-1}(a)$, $d = \alpha^{-1}(b)$, and $\frac{du}{dt} = \frac{1}{\alpha'(u)}$. Also, $dt = \alpha'(u)\,du$, and by the Chain Rule

$$\tilde{x}'(u) = \frac{d\tilde{x}}{du} = \frac{d}{du}(x(\alpha(u))) = \frac{dx}{dt}\frac{dt}{du} = x'(t)\,\alpha'(u) \quad\Rightarrow\quad x'(t) = \frac{\tilde{x}'(u)}{\alpha'(u)}$$

so making the susbstitution $t = \alpha(u)$ gives

$$\begin{aligned}\int_a^b P(x(t),y(t))\,x'(t)\,dt &= \int_{\alpha^{-1}(a)}^{\alpha^{-1}(b)} P(x(\alpha(u)),y(\alpha(u)))\,\frac{\tilde{x}'(u)}{\alpha'(u)}\,(\alpha'(u)\,du)\\ &= \int_c^d P(\tilde{x}(u),\tilde{y}(u))\,\tilde{x}'(u)\,du\ ,\end{aligned}$$

which shows that $\int_C P(x,y)\,dx$ has the same value for both parametrizations. A similar argument shows that $\int_C Q(x,y)\,dy$ has the same value for both parametrizations, and hence $\int_C \mathbf{f}\cdot d\mathbf{r}$ has the same value. **QED**

Notice that the condition $\alpha'(u) > 0$ in Theorem 4.2 means that the two parametrizations move along C in the same direction. That was *not* the case with the "reverse" parametrization for $-C$: for $u = a + b - t$ we have $t = \alpha(u) = a + b - u \Rightarrow \alpha'(u) = -1 < 0$.

Example 4.4. Evaluate the line integral $\int_C (x^2+y^2)\,dx + 2xy\,dy$ from Example 4.2, Section 4.1, along the curve $C: x=t,\ y=2t^2,\ 0 \le t \le 1$, where $t = \sin u$ for $0 \le u \le \pi/2$.

Solution: First, we notice that $0 = \sin 0$, $1 = \sin(\pi/2)$, and $\frac{dt}{du} = \cos u > 0$ on $(0,\pi/2)$. So by Theorem 4.2 we know that if C is parametrized by

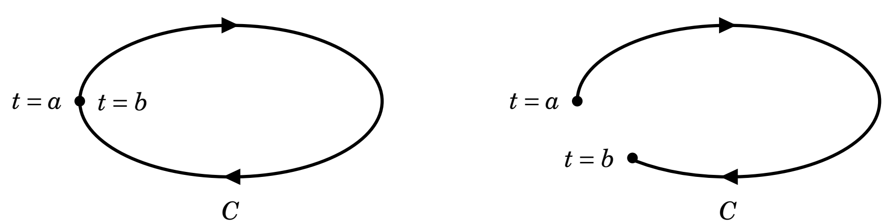
$$x = \sin u\ , \quad y = 2\sin^2 u\ , \quad 0 \le u \le \pi/2$$

then $\int_C (x^2+y^2)\,dx + 2xy\,dy$ should have the same value as we found in Example 4.2, namely $\frac{13}{3}$. And we can indeed verify this:

$$\begin{aligned}
\int_C (x^2+y^2)\,dx + 2xy\,dy &= \int_0^{\pi/2} \Big((\sin^2 u + (2\sin^2 u)^2)\cos u + 2(\sin u)(2\sin^2 u)4\sin u \,\cos u \Big)\,du \\
&= \int_0^{\pi/2} \Big(\sin^2 u + 20\sin^4 u\Big)\cos u\,du \\
&= \frac{\sin^3 u}{3} + 4\sin^5 u \,\Bigg|_0^{\pi/2} \\
&= \frac{1}{3} + 4 \;=\; \frac{13}{3}
\end{aligned}$$

In other words, the line integral is unchanged whether t or u is the parameter for C.

By a **closed curve**, we mean a curve C whose initial point and terminal point are the same, i.e. for $C: x = x(t),\ y = y(t),\ a \le t \le b$, we have $(x(a), y(a)) = (x(b), y(b))$.



(a) Closed **(b)** Not closed

Figure 4.2.1 Closed vs nonclosed curves

A **simple closed curve** is a closed curve which does not intersect itself. Note that any closed curve can be regarded as a union of simple closed curves (think of the loops in a figure eight). We use the special notation

$$\oint_C f(x,y)\,ds \quad \text{and} \quad \oint_C \mathbf{f}\cdot d\mathbf{r}$$

to denote line integrals of scalar and vector fields, respectively, along closed curves. In some older texts you may see the notation $\ointctrclockwise$ or $\ointclockwise$ to indicate a line integral traversing a closed curve in a counterclockwise or clockwise direction, respectively.

So far, the examples we have seen of line integrals (e.g. Example 4.2) have had the same value for different curves joining the initial point to the terminal point. That is, the line integral has been independent of the path joining the two points. As we mentioned before, this is not always the case. The following theorem gives a necessary and sufficient condition for this *path independence*:

Theorem 4.3. In a region R, the line integral $\int_C \mathbf{f} \cdot d\mathbf{r}$ is independent of the path between any two points in R if and only if $\oint_C \mathbf{f} \cdot d\mathbf{r} = 0$ for every closed curve C which is contained in R.

Proof: Suppose that $\oint_C \mathbf{f} \cdot d\mathbf{r} = 0$ for every closed curve C which is contained in R. Let P_1 and P_2 be two distinct points in R. Let C_1 be a curve in R going from P_1 to P_2, and let C_2 be another curve in R going from P_1 to P_2, as in Figure 4.2.2.

Then $C = C_1 \cup -C_2$ is a closed curve in R (from P_1 to P_1), and so $\oint_C \mathbf{f} \cdot d\mathbf{r} = 0$. Thus,

$$\begin{aligned} 0 &= \oint_C \mathbf{f} \cdot d\mathbf{r} \\ &= \int_{C_1} \mathbf{f} \cdot d\mathbf{r} + \int_{-C_2} \mathbf{f} \cdot d\mathbf{r} \\ &= \int_{C_1} \mathbf{f} \cdot d\mathbf{r} - \int_{C_2} \mathbf{f} \cdot d\mathbf{r} \text{, and so} \end{aligned}$$

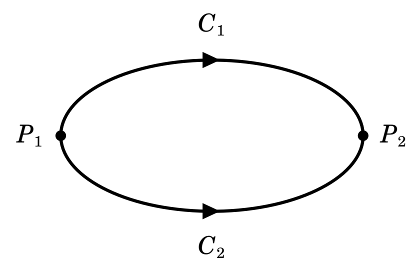


Figure 4.2.2

$\int_{C_1} \mathbf{f} \cdot d\mathbf{r} = \int_{C_2} \mathbf{f} \cdot d\mathbf{r}$. This proves path independence.

Conversely, suppose that the line integral $\int_C \mathbf{f} \cdot d\mathbf{r}$ is independent of the path between any two points in R. Let C be a closed curve contained in R. Let P_1 and P_2 be two distinct points on C. Let C_1 be a part of the curve C that goes from P_1 to P_2, and let C_2 be the remaining part of C that goes from P_1 to P_2, again as in Figure 4.2.2. Then by path independence we have

$$\begin{aligned} \int_{C_1} \mathbf{f} \cdot d\mathbf{r} &= \int_{C_2} \mathbf{f} \cdot d\mathbf{r} \\ \int_{C_1} \mathbf{f} \cdot d\mathbf{r} - \int_{C_2} \mathbf{f} \cdot d\mathbf{r} &= 0 \\ \int_{C_1} \mathbf{f} \cdot d\mathbf{r} + \int_{-C_2} \mathbf{f} \cdot d\mathbf{r} &= 0 \text{, so} \\ \oint_C \mathbf{f} \cdot d\mathbf{r} &= 0 \end{aligned}$$

since $C = C_1 \cup -C_2$. **QED**

Clearly, the above theorem does not give a practical way to determine path independence, since it is impossible to check the line integrals around all possible closed curves in a region. What it mostly does is give an idea of the way in which line integrals behave, and how seemingly unrelated line integrals can be related (in this case, a specific line integral between two points and *all* line integrals around closed curves).

For a more practical method for determining path independence, we first need a version of the Chain Rule for multivariable functions:

Theorem 4.4. (**Chain Rule**) If $z = f(x,y)$ is a continuously differentiable function of x and y, and both $x = x(t)$ and $y = y(t)$ are differentiable functions of t, then z is a differentiable function of t, and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} \tag{4.19}$$

at all points where the derivatives on the right are defined.

The proof is virtually identical to the proof of Theorem 2.2 from Section 2.4 (which uses the Mean Value Theorem), so we omit it.[1] We will now use this Chain Rule to prove the following *sufficient* condition for path independence of line integrals:

Theorem 4.5. Let $\mathbf{f}(x,y) = P(x,y)\mathbf{i} + Q(x,y)\mathbf{j}$ be a vector field in some region R, with P and Q continuously differentiable functions on R. Let C be a smooth curve in R parametrized by $x = x(t)$, $y = y(t)$, $a \le t \le b$. Suppose that there is a real-valued function $F(x,y)$ such that $\nabla F = \mathbf{f}$ on R. Then

$$\int_C \mathbf{f}\cdot d\mathbf{r} = F(B) - F(A)\,, \tag{4.20}$$

where $A = (x(a), y(a))$ and $B = (x(b), y(b))$ are the endpoints of C. Thus, the line integral is independent of the path between its endpoints, since it depends only on the values of F at those endpoints.

Proof: By definition of $\int_C \mathbf{f}\cdot d\mathbf{r}$, we have

$$\begin{aligned}
\int_C \mathbf{f}\cdot d\mathbf{r} &= \int_a^b \left(P(x(t),y(t))x'(t) + Q(x(t),y(t))y'(t)\right) dt \\
&= \int_a^b \left(\frac{\partial F}{\partial x}\frac{dx}{dt} + \frac{\partial F}{\partial y}\frac{dy}{dt}\right) dt \ \ \left(\text{since } \nabla F = \mathbf{f} \Rightarrow \frac{\partial F}{\partial x} = P \text{ and } \frac{\partial F}{\partial y} = Q\right) \\
&= \int_a^b F'(x(t),y(t))\,dt \ \ (\text{by the Chain Rule in Theorem 4.4}) \\
&= F(x(t),y(t))\Big|_a^b = F(B) - F(A)
\end{aligned}$$

by the Fundamental Theorem of Calculus. **QED**

[1] See TAYLOR and MANN, § 6.5.

Theorem 4.5 can be thought of as the line integral version of the Fundamental Theorem of Calculus. A real-valued function $F(x,y)$ such that $\nabla F(x,y) = \mathbf{f}(x,y)$ is called a **potential** for $\mathbf{f}$. A **conservative** vector field is one which has a potential.

Example 4.5. Recall from Examples 4.2 and 4.3 in Section 4.1 that the line integral $\int_C (x^2 + y^2)\,dx + 2xy\,dy$ was found to have the value $\frac{13}{3}$ for three different curves C going from the point $(0,0)$ to the point $(1,2)$. Use Theorem 4.5 to show that this line integral is indeed path independent.

Solution: We need to find a real-valued function $F(x,y)$ such that

$$\frac{\partial F}{\partial x} = x^2 + y^2 \quad \text{and} \quad \frac{\partial F}{\partial y} = 2xy \;.$$

Suppose that $\frac{\partial F}{\partial x} = x^2 + y^2$, Then we must have $F(x,y) = \frac{1}{3}x^3 + xy^2 + g(y)$ for some function $g(y)$. So $\frac{\partial F}{\partial y} = 2xy + g'(y)$ satisfies the condition $\frac{\partial F}{\partial y} = 2xy$ if $g'(y) = 0$, i.e. $g(y) = K$, where K is a constant. Since any choice for K will do (why?), we pick $K = 0$. Thus, a potential $F(x,y)$ for $\mathbf{f}(x,y) = (x^2 + y^2)\,\mathbf{i} + 2xy\,\mathbf{j}$ exists, namely

$$F(x,y) \;=\; \frac{1}{3}x^3 + xy^2 \;.$$

Hence the line integral $\int_C (x^2 + y^2)\,dx + 2xy\,dy$ is path independent.
Note that we can also verify that the value of the line integral of $\mathbf{f}$ along any curve C going from $(0,0)$ to $(1,2)$ will always be $\frac{13}{3}$, since by Theorem 4.5

$$\int_C \mathbf{f}\cdot d\mathbf{r} \;=\; F(1,2) \;-\; F(0,0) \;=\; \frac{1}{3}(1)^3 + (1)(2)^2 - (0+0) \;=\; \frac{1}{3} + 4 \;=\; \frac{13}{3}\;.$$

A consequence of Theorem 4.5 in the special case where C is a closed curve, so that the endpoints A and B are the same point, is the following important corollary:

Corollary 4.6. If a vector field $\mathbf{f}$ has a potential in a region R, then $\oint_C \mathbf{f}\cdot d\mathbf{r} = 0$ for any closed curve C in R (i.e. $\oint_C \nabla F \cdot d\mathbf{r} = 0$ for any real-valued function $F(x,y)$).

Example 4.6. Evaluate $\oint_C x\,dx + y\,dy$ for $C : x = 2\cos t,\; y = 3\sin t,\; 0 \le t \le 2\pi$.

Solution: The vector field $\mathbf{f}(x,y) = x\,\mathbf{i} + y\,\mathbf{j}$ has a potential $F(x,y)$:

$$\frac{\partial F}{\partial x} = x \;\Rightarrow\; F(x,y) \;=\; \frac{1}{2}x^2 + g(y)\;,\text{so}$$
$$\frac{\partial F}{\partial y} = y \;\Rightarrow\; g'(y) = y \;\Rightarrow\; g(y) \;=\; \frac{1}{2}y^2 + K$$

for any constant K, so $F(x,y)=\frac{1}{2}x^2+\frac{1}{2}y^2$ is a potential for $\mathbf{f}(x,y)$. Thus,

$$\oint_C x\,dx+y\,dy \;=\; \oint_C \mathbf{f}\cdot d\mathbf{r} \;=\; 0$$

by Corollary 4.6, since the curve C is closed (it is the ellipse $\frac{x^2}{4}+\frac{y^2}{9}=1$).

Exercises

A

1. Evaluate $\oint_C (x^2+y^2)\,dx+2xy\,dy$ for $C: x=\cos t,\ y=\sin t,\ 0\le t\le 2\pi$.

2. Evaluate $\int_C (x^2+y^2)\,dx+2xy\,dy$ for $C: x=\cos t,\ y=\sin t,\ 0\le t\le \pi$.

3. Is there a potential $F(x,y)$ for $\mathbf{f}(x,y)=y\,\mathbf{i}-x\,\mathbf{j}$? If so, find one.

4. Is there a potential $F(x,y)$ for $\mathbf{f}(x,y)=x\,\mathbf{i}-y\,\mathbf{j}$? If so, find one.

5. Is there a potential $F(x,y)$ for $\mathbf{f}(x,y)=xy^2\,\mathbf{i}+x^3y\,\mathbf{j}$? If so, find one.

B

6. Let $\mathbf{f}(x,y)$ and $\mathbf{g}(x,y)$ be vector fields, let a and b be constants, and let C be a curve in $\mathbb{R}^2$. Show that

$$\int_C (a\,\mathbf{f}\pm b\,\mathbf{g})\cdot d\mathbf{r} \;=\; a\int_C \mathbf{f}\cdot d\mathbf{r} \;\pm\; b\int_C \mathbf{g}\cdot d\mathbf{r}\,.$$

7. Let C be a curve whose arc length is L. Show that $\int_C 1\,ds=L$.

8. Let $f(x,y)$ and $g(x,y)$ be continuously differentiable real-valued functions in a region R. Show that

$$\oint_C f\,\nabla g\cdot d\mathbf{r} \;=\; -\oint_C g\,\nabla f\cdot d\mathbf{r}$$

for any closed curve C in R. (*Hint: Use Exercise 21 in Section 2.4.*)

9. Let $\mathbf{f}(x,y)=\frac{-y}{x^2+y^2}\,\mathbf{i}+\frac{x}{x^2+y^2}\,\mathbf{j}$ for all $(x,y)\ne(0,0)$, and $C: x=\cos t,\ y=\sin t,\ 0\le t\le 2\pi$.

(a) Show that $\mathbf{f}=\nabla F$, for $F(x,y)=\tan^{-1}(y/x)$.

(b) Show that $\oint_C \mathbf{f}\cdot d\mathbf{r}=2\pi$. Does this contradict Corollary 4.6? Explain.

C

10. Let $g(x)$ and $h(y)$ be differentiable functions, and let $\mathbf{f}(x,y)=h(y)\,\mathbf{i}+g(x)\,\mathbf{j}$. Can $\mathbf{f}$ have a potential $F(x,y)$? If so, find it. You may assume that F would be smooth. (*Hint: Consider the mixed partial derivatives of F.*)

4.3 Green's Theorem

We will now see a way of evaluating the line integral of a *smooth* vector field around a simple closed curve. A vector field $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ is **smooth** if its component functions $P(x,y)$ and $Q(x,y)$ are smooth. We will use *Green's Theorem* (sometimes called *Green's Theorem in the plane*) to relate the *line* integral around a closed curve with a *double* integral over the region inside the curve:

Theorem 4.7. **(Green's Theorem)** Let R be a region in $\mathbb{R}^2$ whose boundary is a simple closed curve C which is piecewise smooth. Let $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ be a smooth vector field defined on both R and C. Then

$$\oint_C \mathbf{f}\cdot d\mathbf{r} \;=\; \iint\limits_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA\,, \tag{4.21}$$

where C is traversed so that R is always on the left side of C.

Proof: We will prove the theorem in the case for a *simple* region R, that is, where the boundary curve C can be written as $C = C_1 \cup C_2$ in two distinct ways:

$$C_1 = \text{the curve } y = y_1(x) \text{ from the point } X_1 \text{ to the point } X_2 \tag{4.22}$$
$$C_2 = \text{the curve } y = y_2(x) \text{ from the point } X_2 \text{ to the point } X_1, \tag{4.23}$$

where X_1 and X_2 are the points on C farthest to the left and right, respectively; and

$$C_1 = \text{the curve } x = x_1(y) \text{ from the point } Y_2 \text{ to the point } Y_1 \tag{4.24}$$
$$C_2 = \text{the curve } x = x_2(y) \text{ from the point } Y_1 \text{ to the point } Y_2, \tag{4.25}$$

where Y_1 and Y_2 are the lowest and highest points, respectively, on C. See Figure 4.3.1.

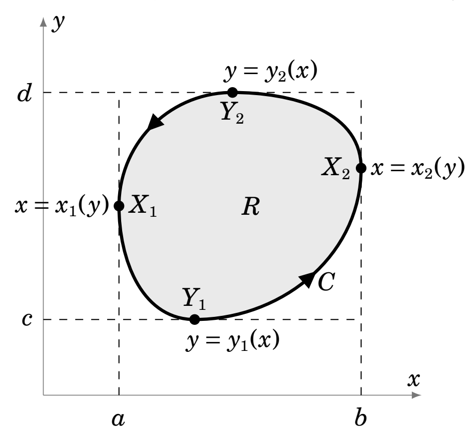


Figure 4.3.1

Integrate $P(x,y)$ around C using the representation $C = C_1 \cup C_2$ given by (4.23) and (4.24).

Since $y = y_1(x)$ along C_1 (as x goes from a to b) and $y = y_2(x)$ along C_2 (as x goes from b to a), as we see from Figure 4.3.1, then we have

$$\begin{aligned}
\oint_C P(x,y)\,dx &= \int_{C_1} P(x,y)\,dx + \int_{C_2} P(x,y)\,dx \\
&= \int_a^b P(x,y_1(x))\,dx + \int_b^a P(x,y_2(x))\,dx \\
&= \int_a^b P(x,y_1(x))\,dx - \int_a^b P(x,y_2(x))\,dx \\
&= -\int_a^b (P(x,y_2(x)) - P(x,y_1(x)))\,dx \\
&= -\int_a^b \left(P(x,y)\Big|_{y=y_1(x)}^{y=y_2(x)} \right) dx \\
&= -\int_a^b \int_{y_1(x)}^{y_2(x)} \frac{\partial P(x,y)}{\partial y}\,dy\,dx \quad \text{(by the Fundamental Theorem of Calculus)} \\
&= -\iint_R \frac{\partial P}{\partial y}\,dA\,.
\end{aligned}$$

Likewise, integrate $Q(x,y)$ around C using the representation $C = C_1 \cup C_2$ given by (4.25) and (4.26). Since $x = x_1(y)$ along C_1 (as y goes from d to c) and $x = x_2(y)$ along C_2 (as y goes from c to d), as we see from Figure 4.3.1, then we have

$$\begin{aligned}
\oint_C Q(x,y)\,dy &= \int_{C_1} Q(x,y)\,dy + \int_{C_2} Q(x,y)\,dy \\
&= \int_d^c Q(x_1(y),y)\,dy + \int_c^d Q(x_2(y),y)\,dy \\
&= -\int_c^d Q(x_1(y),y)\,dy + \int_c^d Q(x_2(y),y)\,dy \\
&= \int_c^d (Q(x_2(y),y) - Q(x_1(y),y))\,dy \\
&= \int_c^d \left(Q(x,y)\Big|_{x=x_1(y)}^{x=x_2(y)} \right) dy \\
&= \int_c^d \int_{x_1(y)}^{x_2(y)} \frac{\partial Q(x,y)}{\partial x}\,dx\,dy \quad \text{(by the Fundamental Theorem of Calculus)} \\
&= \iint_R \frac{\partial Q}{\partial x}\,dA\,, \text{ and so}
\end{aligned}$$

$$\begin{aligned}\oint_C \mathbf{f}\cdot d\mathbf{r} &= \oint_C P(x,y)\,dx + \oint_C Q(x,y)\,dy\\ &= -\iint\limits_R \frac{\partial P}{\partial y}\,dA + \iint\limits_R \frac{\partial Q}{\partial x}\,dA\\ &= \iint\limits_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA\ .\end{aligned}$$

QED

Though we proved Green's Theorem only for a simple region R, the theorem can also be proved for more general regions (say, a union of simple regions).[2]

Example 4.7. Evaluate $\oint_C (x^2+y^2)\,dx + 2xy\,dy$, where C is the boundary (traversed counterclockwise) of the region $R = \{(x,y) : 0 \le x \le 1,\ 2x^2 \le y \le 2x\}$.

Solution: R is the shaded region in Figure 4.3.2. By Green's Theorem, for $P(x,y) = x^2+y^2$ and $Q(x,y) = 2xy$, we have

$$\begin{aligned}\oint_C (x^2+y^2)\,dx + 2xy\,dy &= \iint\limits_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA\\ &= \iint\limits_R (2y-2y)\,dA \;=\; \iint\limits_R 0\,dA \;=\; 0\ .\end{aligned}$$

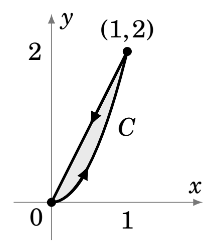


Figure 4.3.2

We actually already knew that the answer was zero. Recall from Example 4.5 in Section 4.2 that the vector field $\mathbf{f}(x,y) = (x^2+y^2)\,\mathbf{i} + 2xy\,\mathbf{j}$ has a potential function $F(x,y) = \frac{1}{3}x^3 + xy^2$, and so $\oint_C \mathbf{f}\cdot d\mathbf{r} = 0$ by Corollary 4.6.

Example 4.8. Let $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$, where

$$P(x,y) = \frac{-y}{x^2+y^2} \quad \text{and} \quad Q(x,y) = \frac{x}{x^2+y^2}\ ,$$

and let $R = \{(x,y) : 0 < x^2+y^2 \le 1\}$. For the boundary curve $C : x^2+y^2 = 1$, traversed counterclockwise, it was shown in Exercise 9(b) in Section 4.2 that $\oint_C \mathbf{f}\cdot d\mathbf{r} = 2\pi$. But

$$\frac{\partial Q}{\partial x} = \frac{y^2-x^2}{(x^2+y^2)^2} = \frac{\partial P}{\partial y} \;\Rightarrow\; \iint\limits_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \;=\; \iint\limits_R 0\,dA = 0\ .$$

This would seem to contradict Green's Theorem. However, note that R is not the *entire* region enclosed by C, since the point $(0,0)$ is not contained in R. That is, R has a "hole" at the origin, so Green's Theorem does not apply.

[2] See TAYLOR and MANN, § 15.31 for a discussion of some of the difficulties involved when the boundary curve is "complicated".

If we modify the region R to be the *annulus* $R = \{(x, y) : 1/4 \le x^2 + y^2 \le 1\}$ (see Figure 4.3.3), and take the "boundary" C of R to be $C = C_1 \cup C_2$, where C_1 is the unit circle $x^2 + y^2 = 1$ traversed counterclockwise and C_2 is the circle $x^2 + y^2 = 1/4$ traversed *clockwise*, then it can be shown (see Exercise 8) that

$$\oint_C \mathbf{f} \cdot d\mathbf{r} = 0 \,.$$

We would still have $\iint\limits_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = 0$, so for this R we would have

$$\oint_C \mathbf{f} \cdot d\mathbf{r} = \iint\limits_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \,,$$

which shows that Green's Theorem holds for the annular region R.

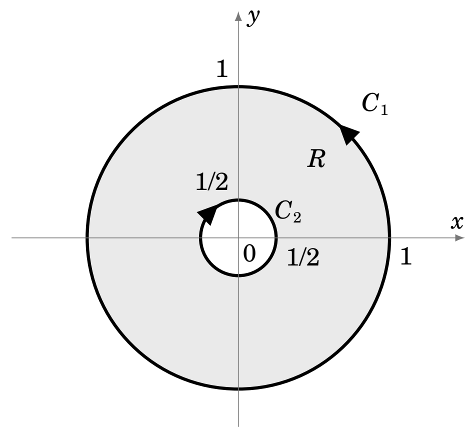
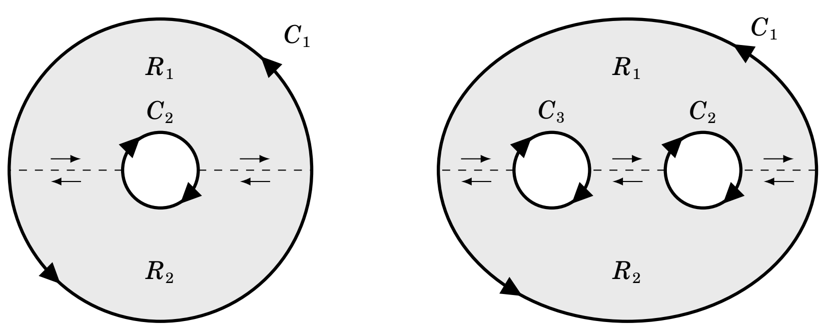


Figure 4.3.3 The annulus R

It turns out that Green's Theorem can be extended to *multiply connected* regions, that is, regions like the annulus in Example 4.8, which have one or more *regions* cut out from the interior, as opposed to discrete *points* being cut out. For such regions, the "outer" boundary and the "inner" boundaries are traversed so that R is always on the left side.



(a) Region R with one hole **(b)** Region R with two holes

Figure 4.3.4 Multiply connected regions

The intuitive idea for why Green's Theorem holds for multiply connected regions is shown in Figure 4.3.4 above. The idea is to cut "slits" between the boundaries of a multiply connected region R so that R is divided into subregions which do *not* have any "holes". For example, in Figure 4.3.4(a) the region R is the union of the regions R_1 and R_2, which are divided by the slits indicated by the dashed lines. Those slits are part of the boundary of both R_1 and R_2, and we traverse then in the manner indicated by the arrows. Notice that along each slit the boundary of R_1 is traversed in the opposite direction as that of R_2, which

means that the line integrals of $\mathbf{f}$ along those slits cancel each other out. Since R_1 and R_2 do not have holes in them, then Green's Theorem holds in each subregion, so that

$$\oint_{\substack{\text{bdy}\\ \text{of } R_1}} \mathbf{f}\cdot d\mathbf{r} = \iint\limits_{R_1} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \quad \text{and} \quad \oint_{\substack{\text{bdy}\\ \text{of } R_2}} \mathbf{f}\cdot d\mathbf{r} = \iint\limits_{R_2} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \;.$$

But since the line integrals along the slits cancel out, we have

$$\oint_{C_1 \cup C_2} \mathbf{f}\cdot d\mathbf{r} = \oint_{\substack{\text{bdy}\\ \text{of } R_1}} \mathbf{f}\cdot d\mathbf{r} + \oint_{\substack{\text{bdy}\\ \text{of } R_2}} \mathbf{f}\cdot d\mathbf{r} \;,$$

and so

$$\oint_{C_1 \cup C_2} \mathbf{f}\cdot d\mathbf{r} = \iint\limits_{R_1} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA + \iint\limits_{R_2} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \iint\limits_{R} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \;,$$

which shows that Green's Theorem holds in the region R. A similar argument shows that the theorem holds in the region with two holes shown in Figure 4.3.4(b).

We know from Corollary 4.6 that when a smooth vector field $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ on a region R (whose boundary is a piecewise smooth, simple closed curve C) has a potential in R, then $\oint_C \mathbf{f}\cdot d\mathbf{r} = 0$. And if the potential $F(x,y)$ is smooth in R, then $\frac{\partial F}{\partial x} = P$ and $\frac{\partial F}{\partial y} = Q$, and so we know that

$$\frac{\partial^2 F}{\partial y\,\partial x} = \frac{\partial^2 F}{\partial x\,\partial y} \;\Rightarrow\; \frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \quad \text{in } R.$$

Conversely, if $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$ in R then

$$\oint_C \mathbf{f}\cdot d\mathbf{r} = \iint\limits_{R} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \iint\limits_{R} 0\,dA = 0 \;.$$

For a **simply connected** region R (i.e. a region with no holes), the following can be shown:

The following statements are equivalent for a simply connected region R in $\mathbb{R}^2$:

(a) $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ has a smooth potential $F(x,y)$ in R

(b) $\displaystyle\int_C \mathbf{f}\cdot d\mathbf{r}$ is independent of the path for any curve C in R

(c) $\displaystyle\oint_C \mathbf{f}\cdot d\mathbf{r} = 0$ for every simple closed curve C in R

(d) $\dfrac{\partial P}{\partial y} = \dfrac{\partial Q}{\partial x}$ in R (in this case, the differential form $P\,dx + Q\,dy$ is exact)

Exercises

A

For Exercises 1-4, use Green's Theorem to evaluate the given line integral around the curve C, traversed counterclockwise.

1. $\oint_C (x^2 - y^2)\,dx + 2xy\,dy$; C is the boundary of $R = \{(x,y) : 0 \le x \le 1,\ 2x^2 \le y \le 2x\}$

2. $\oint_C x^2 y\,dx + 2xy\,dy$; C is the boundary of $R = \{(x,y) : 0 \le x \le 1,\ x^2 \le y \le x\}$

3. $\oint_C 2y\,dx - 3x\,dy$; C is the circle $x^2 + y^2 = 1$

4. $\oint_C (e^{x^2} + y^2)\,dx + (e^{y^2} + x^2)\,dy$; C is the boundary of the triangle with vertices $(0,0)$, $(4,0)$ and $(0,4)$

5. Is there a potential $F(x,y)$ for $\mathbf{f}(x,y) = (y^2 + 3x^2)\,\mathbf{i} + 2xy\,\mathbf{j}$? If so, find one.

6. Is there a potential $F(x,y)$ for $\mathbf{f}(x,y) = (x^3\cos(xy) + 2x\sin(xy))\,\mathbf{i} + x^2 y\cos(xy)\,\mathbf{j}$? If so, find one.

7. Is there a potential $F(x,y)$ for $\mathbf{f}(x,y) = (8xy + 3)\,\mathbf{i} + 4(x^2 + y)\,\mathbf{j}$? If so, find one.

8. Show that for any constants a, b and any closed simple curve C, $\oint_C a\,dx + b\,dy = 0$.

B

9. For the vector field $\mathbf{f}$ as in Example 4.8, show directly that $\oint_C \mathbf{f} \cdot d\mathbf{r} = 0$, where C is the boundary of the annulus $R = \{(x,y) : 1/4 \le x^2 + y^2 \le 1\}$ traversed so that R is always on the left.

10. Evaluate $\oint_C e^x \sin y\,dx + (y^3 + e^x \cos y)\,dy$, where C is the boundary of the rectangle with vertices $(1,-1)$, $(1,1)$, $(-1,1)$ and $(-1,-1)$, traversed counterclockwise.

C

11. For a region R bounded by a simple closed curve C, show that the area A of R is

$$A = -\oint_C y\,dx = \oint_C x\,dy = \frac{1}{2}\oint_C x\,dy - y\,dx\,,$$

where C is traversed so that R is always on the left. (*Hint: Use Green's Theorem and the fact that* $A = \iint\limits_R 1\,dA$.)

4.4 Surface Integrals and the Divergence Theorem

In Section 4.1 we learned how to integrate along a curve. We will now learn how to perform integration over a *surface* in $\mathbb{R}^3$, such as a sphere or a paraboloid. Recall from Section 1.8 how we identified points (x, y, z) on a curve C in $\mathbb{R}^3$, parametrized by $x = x(t)$, $y = y(t)$, $z = z(t)$, $a \le t \le b$, with the terminal points of the position vector

$$\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \quad \text{for } t \text{ in } [a, b].$$

The idea behind a parametrization of a curve is that it "transforms" a subset of $\mathbb{R}^1$ (normally an interval $[a, b]$) into a curve in $\mathbb{R}^2$ or $\mathbb{R}^3$ (see Figure 4.4.1).

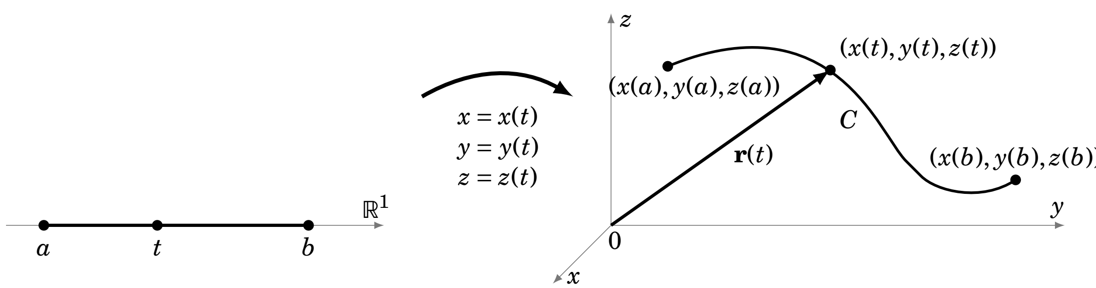


Figure 4.4.1 Parametrization of a curve C in $\mathbb{R}^3$

Similar to how we used a parametrization of a curve to define the line integral along the curve, we will use a parametrization of a surface to define a *surface integral*. We will use *two* variables, u and v, to parametrize a surface Σ in $\mathbb{R}^3$: $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$, for (u, v) in some region R in $\mathbb{R}^2$ (see Figure 4.4.2).

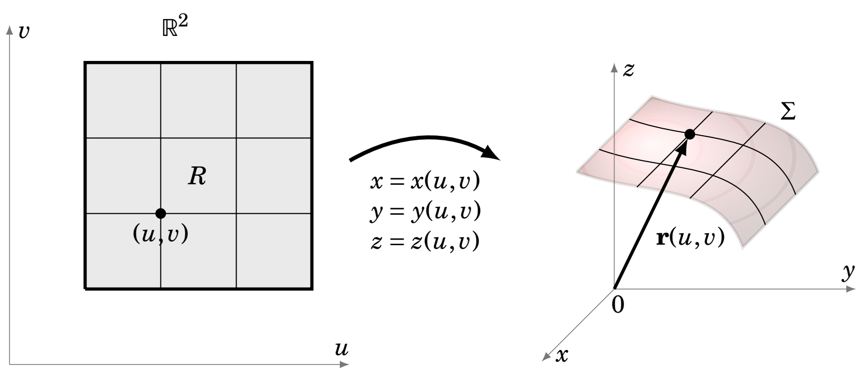


Figure 4.4.2 Parametrization of a surface Σ in $\mathbb{R}^3$

In this case, the position vector of a point on the surface Σ is given by the vector-valued function

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k} \quad \text{for } (u, v) \text{ in } R.$$

Since $\mathbf{r}(u,v)$ is a function of two variables, define the partial derivatives $\frac{\partial \mathbf{r}}{\partial u}$ and $\frac{\partial \mathbf{r}}{\partial v}$ for (u,v) in R by

$$\begin{aligned}\frac{\partial \mathbf{r}}{\partial u}(u,v) &= \frac{\partial x}{\partial u}(u,v)\mathbf{i} + \frac{\partial y}{\partial u}(u,v)\mathbf{j} + \frac{\partial z}{\partial u}(u,v)\mathbf{k}\,, \text{ and}\\ \frac{\partial \mathbf{r}}{\partial v}(u,v) &= \frac{\partial x}{\partial v}(u,v)\mathbf{i} + \frac{\partial y}{\partial v}(u,v)\mathbf{j} + \frac{\partial z}{\partial v}(u,v)\mathbf{k}\,.\end{aligned}$$

The parametrization of Σ can be thought of as "transforming" a region in $\mathbb{R}^2$ (in the uv-plane) into a 2-dimensional surface in $\mathbb{R}^3$. This parametrization of the surface is sometimes called a *patch*, based on the idea of "patching" the region R onto Σ in the grid-like manner shown in Figure 4.4.2.

In fact, those gridlines in R lead us to how we will define a surface integral over Σ. Along the vertical gridlines in R, the variable u is constant. So those lines get mapped to curves on Σ, and the variable u is constant along the position vector $\mathbf{r}(u,v)$. Thus, the tangent vector to those curves at a point (u,v) is $\frac{\partial \mathbf{r}}{\partial v}$. Similarly, the horizontal gridlines in R get mapped to curves on Σ whose tangent vectors are $\frac{\partial \mathbf{r}}{\partial u}$.

Now take a point (u,v) in R as, say, the lower left corner of one of the rectangular grid sections in R, as shown in Figure 4.4.2. Suppose that this rectangle has a small width and height of Δu and Δv, respectively. The corner points of that rectangle are (u,v), $(u+\Delta u,v)$, $(u+\Delta u,v+\Delta v)$ and $(u,v+\Delta v)$. So the area of that rectangle is $A = \Delta u\,\Delta v$. Then that rectangle gets mapped by the parametrization onto some section of the surface Σ which, for Δu and Δv small enough, will have a surface area (call it $d\sigma$) that is very close to the area of the parallelogram which has adjacent sides $\mathbf{r}(u+\Delta u,v)-\mathbf{r}(u,v)$ (corresponding to the line segment from (u,v) to $(u+\Delta u,v)$ in R) and $\mathbf{r}(u,v+\Delta v)-\mathbf{r}(u,v)$ (corresponding to the line segment from (u,v) to $(u,v+\Delta v)$ in R). But by combining our usual notion of a partial derivative (see Definition 2.3 in Section 2.2) with that of the derivative of a vector-valued function (see Definition 1.12 in Section 1.8) applied to a function of two variables, we have

$$\begin{aligned}\frac{\partial \mathbf{r}}{\partial u} &\approx \frac{\mathbf{r}(u+\Delta u,v)-\mathbf{r}(u,v)}{\Delta u}\,, \text{ and}\\ \frac{\partial \mathbf{r}}{\partial v} &\approx \frac{\mathbf{r}(u,v+\Delta v)-\mathbf{r}(u,v)}{\Delta v}\,,\end{aligned}$$

and so the surface area element $d\sigma$ is approximately

$$\left\|(\mathbf{r}(u+\Delta u,v)-\mathbf{r}(u,v))\times(\mathbf{r}(u,v+\Delta v)-\mathbf{r}(u,v))\right\| \approx \left\|(\Delta u\frac{\partial \mathbf{r}}{\partial u})\times(\Delta v\frac{\partial \mathbf{r}}{\partial v})\right\| = \left\|\frac{\partial \mathbf{r}}{\partial u}\times\frac{\partial \mathbf{r}}{\partial v}\right\|\Delta u\,\Delta v$$

by Theorem 1.13 in Section 1.4. Thus, the total surface area S of Σ is approximately the sum of all the quantities $\left\|\frac{\partial \mathbf{r}}{\partial u}\times\frac{\partial \mathbf{r}}{\partial v}\right\|\Delta u\,\Delta v$, summed over the rectangles in R. Taking the limit of that sum as the diagonal of the largest rectangle goes to 0 gives

$$S = \iint_R \left\|\frac{\partial \mathbf{r}}{\partial u}\times\frac{\partial \mathbf{r}}{\partial v}\right\| du\,dv\,. \tag{4.26}$$

We will write the double integral on the right using the special notation

$$\iint_{\Sigma} d\sigma \;=\; \iint_{R} \left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\| du\,dv\ . \tag{4.27}$$

This is a special case of a *surface integral* over the surface Σ, where the surface area element $d\sigma$ can be thought of as $1\,d\sigma$. Replacing 1 by a general real-valued function $f(x,y,z)$ defined in $\mathbb{R}^3$, we have the following:

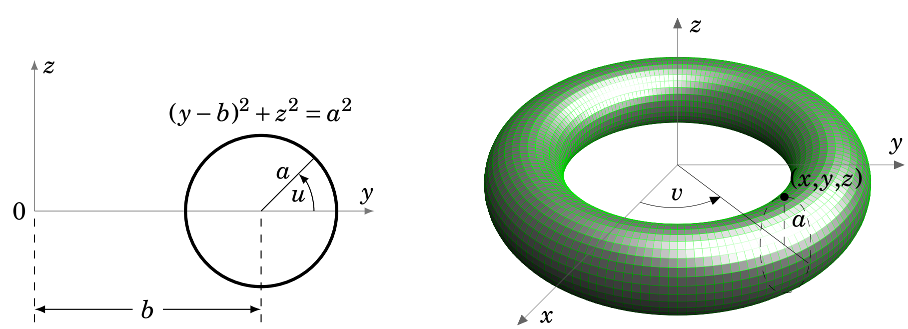
Definition 4.3. Let Σ be a surface in $\mathbb{R}^3$ parametrized by $x = x(u,v)$, $y = y(u,v)$, $z = z(u,v)$, for (u,v) in some region R in $\mathbb{R}^2$. Let $\mathbf{r}(u,v) = x(u,v)\mathbf{i} + y(u,v)\mathbf{j} + z(u,v)\mathbf{k}$ be the position vector for any point on Σ, and let $f(x,y,z)$ be a real-valued function defined on some subset of $\mathbb{R}^3$ that contains Σ. The **surface integral** of $f(x,y,z)$ over Σ is

$$\iint_{\Sigma} f(x,y,z)\,d\sigma \;=\; \iint_{R} f(x(u,v),y(u,v),z(u,v)) \left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\| du\,dv\ . \tag{4.28}$$

In particular, the surface area S of Σ is

$$S \;=\; \iint_{\Sigma} 1\,d\sigma\ . \tag{4.29}$$

Example 4.9. A *torus* T is a surface obtained by revolving a circle of radius a in the yz-plane around the z-axis, where the circle's center is at a distance b from the z-axis ($0 < a < b$), as in Figure 4.4.3. Find the surface area of T.



(a) Circle in the yz-plane **(b)** Torus T

Figure 4.4.3

Solution: For any point on the circle, the line segment from the center of the circle to that point makes an angle u with the y-axis in the positive y direction (see Figure 4.4.3(a)). And as the circle revolves around the z-axis, the line segment from the origin to the center of that

circle sweeps out an angle v with the positive x-axis (see Figure 4.4.3(b)). Thus, the torus can be parametrized as:

$$x = (b + a\cos u)\cos v\,, \quad y = (b + a\cos u)\sin v\,, \quad z = a\sin u\,, \quad 0 \le u \le 2\pi\,, \quad 0 \le v \le 2\pi$$

So for the position vector

$$\begin{aligned}\mathbf{r}(u,v) &= x(u,v)\mathbf{i} + y(u,v)\mathbf{j} + z(u,v)\mathbf{k} \\ &= (b + a\cos u)\cos v\,\mathbf{i} + (b + a\cos u)\sin v\,\mathbf{j} + a\sin u\,\mathbf{k}\end{aligned}$$

we see that

$$\begin{aligned}\frac{\partial \mathbf{r}}{\partial u} &= -a\sin u\,\cos v\,\mathbf{i} - a\sin u\,\sin v\,\mathbf{j} + a\cos u\,\mathbf{k} \\ \frac{\partial \mathbf{r}}{\partial v} &= -(b + a\cos u)\sin v\,\mathbf{i} + (b + a\cos u)\cos v\,\mathbf{j} + 0\mathbf{k}\,,\end{aligned}$$

and so computing the cross product gives

$$\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} = -a(b + a\cos u)\cos v\,\cos u\,\mathbf{i} - a(b + a\cos u)\sin v\,\cos u\,\mathbf{j} - a(b + a\cos u)\sin u\,\mathbf{k}\,,$$

which has magnitude

$$\left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\| = a(b + a\cos u)\,.$$

Thus, the surface area of T is

$$\begin{aligned}S &= \iint_{\Sigma} 1\,d\sigma \\ &= \int_0^{2\pi}\int_0^{2\pi} \left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\| du\,dv \\ &= \int_0^{2\pi}\int_0^{2\pi} a(b + a\cos u)\,du\,dv \\ &= \int_0^{2\pi} \left(abu + a^2 \sin u \Big|_{u=0}^{u=2\pi} \right) dv \\ &= \int_0^{2\pi} 2\pi ab\,dv \\ &= 4\pi^2 ab\end{aligned}$$

Since $\frac{\partial \mathbf{r}}{\partial u}$ and $\frac{\partial \mathbf{r}}{\partial v}$ are tangent to the surface Σ (i.e. lie in the tangent plane to Σ at each point on Σ), then their cross product $\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$ is perpendicular to the tangent plane to the surface at each point of Σ. Thus,

$$\iint_{\Sigma} f(x,y,z)\,d\sigma = \iint_{R} f(x(u,v),y(u,v),z(u,v))\,\|\mathbf{n}\|\,d\sigma\ ,$$

where $\mathbf{n} = \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$. We say that $\mathbf{n}$ is a **normal vector** to Σ.

Recall that normal vectors to a plane can point in two opposite directions. By an **outward unit normal vector** to a surface Σ, we will mean the unit vector that is normal to Σ and points away from the "top" (or "outer" part) of the surface. This is a hazy definition, but the picture in Figure 4.4.4 gives a better idea of what outward normal vectors look like, in the case of a sphere. With this idea in mind, we make the following definition of a surface integral of a 3-dimensional *vector field* over a surface:

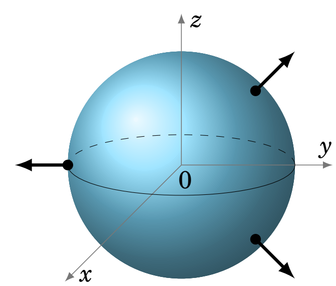


Figure 4.4.4

Definition 4.4. Let Σ be a surface in $\mathbb{R}^3$ and let $\mathbf{f}(x,y,z) = f_1(x,y,z)\mathbf{i} + f_2(x,y,z)\mathbf{j} + f_3(x,y,z)\mathbf{k}$ be a vector field defined on some subset of $\mathbb{R}^3$ that contains Σ. The **surface integral** of $\mathbf{f}$ over Σ is

$$\iint_{\Sigma} \mathbf{f}\cdot d\boldsymbol{\sigma} \ = \ \iint_{\Sigma} \mathbf{f}\cdot\mathbf{n}\,d\sigma\ , \tag{4.30}$$

where, at any point on Σ, $\mathbf{n}$ is the outward unit normal vector to Σ.

Note in the above definition that the dot product inside the integral on the right is a real-valued function, and hence we can use Definition 4.3 to evaluate the integral.

Example 4.10. Evaluate the surface integral $\iint_{\Sigma} \mathbf{f}\cdot d\boldsymbol{\sigma}$, where $\mathbf{f}(x,y,z) = yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$ and Σ is the part of the plane $x+y+z=1$ with $x \ge 0$, $y \ge 0$, and $z \ge 0$, with the outward unit normal $\mathbf{n}$ pointing in the positive z direction (see Figure 4.4.5).

Solution: Since the vector $\mathbf{v} = (1,1,1)$ is normal to the plane $x+y+z=1$ (why?), then dividing $\mathbf{v}$ by its length yields the outward unit normal vector $\mathbf{n} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$. We now need to parametrize Σ. As we can see from Figure 4.4.5, projecting Σ onto the xy-plane yields a triangular region $R = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le 1-x\}$. Thus, using (u,v) instead of (x,y), we see that

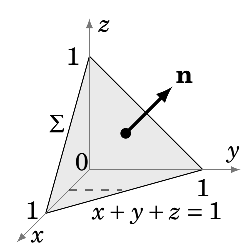


Figure 4.4.5

$$x = u,\ y = v,\ z = 1-(u+v),\quad \text{for } 0 \le u \le 1, 0 \le v \le 1-u$$

is a parametrization of Σ over R (since $z = 1-(x+y)$ on Σ). So on Σ,

$$\begin{aligned}
\mathbf{f}\cdot\mathbf{n} &= (yz,xz,xy)\cdot\left(\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}}\right) = \frac{1}{\sqrt{3}}(yz+xz+xy)\\
&= \frac{1}{\sqrt{3}}((x+y)z+xy) = \frac{1}{\sqrt{3}}((u+v)(1-(u+v))+uv)\\
&= \frac{1}{\sqrt{3}}((u+v)-(u+v)^2+uv)
\end{aligned}$$

for (u,v) in R, and for $\mathbf{r}(u,v) = x(u,v)\mathbf{i}+y(u,v)\mathbf{j}+z(u,v)\mathbf{k} = u\mathbf{i}+v\mathbf{j}+(1-(u+v))\mathbf{k}$ we have

$$\frac{\partial\mathbf{r}}{\partial u}\times\frac{\partial\mathbf{r}}{\partial v} = (1,0,-1)\times(0,1,-1) = (1,1,1) \quad\Rightarrow\quad \left\|\frac{\partial\mathbf{r}}{\partial u}\times\frac{\partial\mathbf{r}}{\partial v}\right\| = \sqrt{3}\,.$$

Thus, integrating over R using vertical slices (e.g. as indicated by the dashed line in Figure 4.4.5) gives

$$\begin{aligned}
\iint_\Sigma \mathbf{f}\cdot d\boldsymbol{\sigma} &= \iint_\Sigma \mathbf{f}\cdot\mathbf{n}\,d\sigma\\
&= \iint_R (\mathbf{f}(x(u,v),y(u,v),z(u,v))\cdot\mathbf{n})\left\|\frac{\partial\mathbf{r}}{\partial u}\times\frac{\partial\mathbf{r}}{\partial v}\right\| dv\,du\\
&= \int_0^1\int_0^{1-u}\frac{1}{\sqrt{3}}((u+v)-(u+v)^2+uv)\sqrt{3}\,dv\,du\\
&= \int_0^1\left(\frac{(u+v)^2}{2}-\frac{(u+v)^3}{3}+\frac{uv^2}{2}\Bigg|_{v=0}^{v=1-u}\right)du\\
&= \int_0^1\left(\frac{1}{6}+\frac{u}{2}-\frac{3u^2}{2}+\frac{5u^3}{6}\right)du\\
&= \frac{u}{6}+\frac{u^2}{4}-\frac{u^3}{2}+\frac{5u^4}{24}\Bigg|_0^1 = \frac{1}{8}\,.
\end{aligned}$$

Computing surface integrals can often be tedious, especially when the formula for the outward unit normal vector at each point of Σ changes. The following theorem provides an easier way in the case when Σ is a **closed surface**, that is, when Σ encloses a bounded solid in $\mathbb{R}^3$. For example, spheres, cubes, and ellipsoids are closed surfaces, but planes and paraboloids are not.

Theorem 4.8. (**Divergence Theorem**) Let Σ be a closed surface in $\mathbb{R}^3$ which bounds a solid S, and let $\mathbf{f}(x, y, z) = f_1(x, y, z)\mathbf{i} + f_2(x, y, z)\mathbf{j} + f_3(x, y, z)\mathbf{k}$ be a vector field defined on some subset of $\mathbb{R}^3$ that contains Σ. Then

$$\iint\limits_{\Sigma} \mathbf{f} \cdot d\boldsymbol{\sigma} = \iiint\limits_{S} \text{div}\, \mathbf{f}\, dV , \tag{4.31}$$

where

$$\text{div}\, \mathbf{f} = \frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y} + \frac{\partial f_3}{\partial z} \tag{4.32}$$

is called the **divergence** of **f**.

The proof of the Divergence Theorem is very similar to the proof of Green's Theorem, i.e. it is first proved for the simple case when the solid S is bounded above by one surface, bounded below by another surface, and bounded laterally by one or more surfaces. The proof can then be extended to more general solids.[3]

Example 4.11. Evaluate $\iint\limits_{\Sigma} \mathbf{f} \cdot d\boldsymbol{\sigma}$, where $\mathbf{f}(x, y, z) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and Σ is the unit sphere $x^2 + y^2 + z^2 = 1$.

Solution: We see that div $\mathbf{f} = 1 + 1 + 1 = 3$, so

$$\begin{aligned}\iint\limits_{\Sigma} \mathbf{f} \cdot d\boldsymbol{\sigma} &= \iiint\limits_{S} \text{div}\, \mathbf{f}\, dV = \iiint\limits_{S} 3\, dV \\ &= 3 \iiint\limits_{S} 1\, dV = 3\,\text{vol}(S) = 3 \cdot \frac{4\pi(1)^3}{3} = 4\pi .\end{aligned}$$

In physical applications, the surface integral $\iint\limits_{\Sigma} \mathbf{f} \cdot d\boldsymbol{\sigma}$ is often referred to as the **flux** of **f** through the surface Σ. For example, if **f** represents the velocity field of a fluid, then the flux is the net quantity of fluid to flow through the surface Σ per unit time. A positive flux means there is a net flow *out* of the surface (i.e. in the direction of the outward unit normal vector **n**), while a negative flux indicates a net flow inward (in the direction of $-\mathbf{n}$).

The term divergence comes from interpreting div **f** as a measure of how much a vector field "diverges" from a point. This is best seen by using another definition of div **f** which is equivalent[4] to the definition given by formula (4.32). Namely, for a point (x, y, z) in $\mathbb{R}^3$,

$$\text{div}\, \mathbf{f}(x, y, z) = \lim_{V \to 0} \frac{1}{V} \iint\limits_{\Sigma} \mathbf{f} \cdot d\boldsymbol{\sigma} , \tag{4.33}$$

[3] See TAYLOR and MANN, § 15.6 for the details.

[4] See SCHEY, p. 36-39, for an intuitive discussion of this.

where V is the volume enclosed by a closed surface Σ around the point (x,y,z). In the limit, $V \to 0$ means that we take smaller and smaller closed surfaces around (x,y,z), which means that the volumes they enclose are going to zero. It can be shown that this limit is independent of the shapes of those surfaces. Notice that the limit being taken is of the ratio of the flux through a surface to the volume enclosed by that surface, which gives a rough measure of the flow "leaving" a point, as we mentioned. Vector fields which have zero divergence are often called *solenoidal* fields.

The following theorem is a simple consequence of formula (4.33).

Theorem 4.9. If the flux of a vector field $\mathbf{f}$ is zero through every closed surface containing a given point, then div $\mathbf{f} = 0$ at that point.

Proof: By formula (4.33), at the given point (x,y,z) we have

$$\begin{aligned} \text{div}\,\mathbf{f}(x,y,z) &= \lim_{V\to 0} \frac{1}{V} \iint_{\Sigma} \mathbf{f}\cdot d\boldsymbol{\sigma} \quad \text{for closed surfaces } \Sigma \text{ containing } (x,y,z)\text{, so} \\ &= \lim_{V\to 0} \frac{1}{V}\,(0) \quad \text{by our assumption that the flux through each } \Sigma \text{ is zero, so} \\ &= \lim_{V\to 0} 0 \\ &= 0\,. \end{aligned}$$

QED

Lastly, we note that sometimes the notation

$$\oiint_{\Sigma} f(x,y,z)\,d\sigma \quad \text{and} \quad \oiint_{\Sigma} \mathbf{f}\cdot d\boldsymbol{\sigma}$$

is used to denote surface integrals of scalar and vector fields, respectively, over closed surfaces. Especially in physics texts, it is common to see simply $\oint_{\Sigma}$ instead of $\oiint_{\Sigma}$.

Exercises

A

For Exercises 1-4, use the Divergence Theorem to evaluate the surface integral $\iint_{\Sigma} \mathbf{f}\cdot d\boldsymbol{\sigma}$ of the given vector field $\mathbf{f}(x,y,z)$ over the surface Σ.

1. $\mathbf{f}(x,y,z) = x\mathbf{i} + 2y\mathbf{j} + 3z\mathbf{k}$, $\Sigma : x^2 + y^2 + z^2 = 9$

2. $\mathbf{f}(x,y,z) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, Σ : boundary of the solid cube $S = \{(x,y,z) : 0 \le x,y,z \le 1\}$

3. $\mathbf{f}(x,y,z) = x^3\mathbf{i} + y^3\mathbf{j} + z^3\mathbf{k}$, $\Sigma : x^2 + y^2 + z^2 = 1$

4. $\mathbf{f}(x,y,z) = 2\mathbf{i} + 3\mathbf{j} + 5\mathbf{k}$, $\Sigma : x^2 + y^2 + z^2 = 1$

B

5. Show that the flux of any constant vector field through any closed surface is zero.

6. Evaluate the surface integral from Exercise 2 *without* using the Divergence Theorem, i.e. using only Definition 4.3, as in Example 4.10. Note that there will be a different outward unit normal vector to each of the six faces of the cube.

7. Evaluate the surface integral $\iint_{\Sigma} \mathbf{f} \cdot d\boldsymbol{\sigma}$, where $\mathbf{f}(x, y, z) = x^2\mathbf{i} + xy\mathbf{j} + z\mathbf{k}$ and Σ is the part of the plane $6x + 3y + 2z = 6$ with $x \ge 0$, $y \ge 0$, and $z \ge 0$, with the outward unit normal $\mathbf{n}$ pointing in the positive z direction.

8. Use a surface integral to show that the surface area of a sphere of radius r is $4\pi r^2$. (*Hint: Use spherical coordinates to parametrize the sphere.*)

9. Use a surface integral to show that the surface area of a right circular cone of radius R and height h is $\pi R\sqrt{h^2 + R^2}$. (*Hint: Use the parametrization* $x = r\cos\theta$, $y = r\sin\theta$, $z = \frac{h}{R}r$, *for* $0 \le r \le R$ *and* $0 \le \theta \le 2\pi$.)

10. The ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ can be parametrized using *ellipsoidal coordinates*

$$x = a\sin\phi\cos\theta\ ,\ y = b\sin\phi\sin\theta\ ,\ z = c\cos\phi\ ,\ \text{ for } 0 \le \theta \le 2\pi \text{ and } 0 \le \phi \le \pi.$$

Show that the surface area S of the ellipsoid is

$$S = \int_0^{\pi}\int_0^{2\pi} \sin\phi\sqrt{a^2b^2\cos^2\phi + c^2(a^2\sin^2\theta + b^2\cos^2\theta)\sin^2\phi}\, d\theta\, d\phi\ .$$

(Note: The above double integral can not be evaluated by elementary means. For specific values of a, b and c it can be evaluated using numerical methods. An alternative is to express the surface area in terms of *elliptic integrals*.[5])

C

11. Use Definition 4.3 to prove that the surface area S over a region R in $\mathbb{R}^2$ of a surface $z = f(x, y)$ is given by the formula

$$S = \iint_R \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dA\ .$$

(*Hint: Think of the parametrization of the surface.*)

[5] Bowman, F., *Introduction to Elliptic Functions, with Applications*, New York: Dover, 1961, § III.7.

4.5 Stokes' Theorem

So far the only types of line integrals which we have discussed are those along curves in $\mathbb{R}^2$. But the definitions and properties which were covered in Sections 4.1 and 4.2 can easily be extended to include functions of three variables, so that we can now discuss line integrals along curves in $\mathbb{R}^3$.

Definition 4.5. For a real-valued function $f(x,y,z)$ and a curve C in $\mathbb{R}^3$, parametrized by $x = x(t)$, $y = y(t)$, $z = z(t)$, $a \le t \le b$, the **line integral of $f(x,y,z)$ along C with respect to arc length** s is

$$\int_C f(x,y,z)\,ds = \int_a^b f(x(t),y(t),z(t))\sqrt{x'(t)^2+y'(t)^2+z'(t)^2}\,dt\,. \tag{4.34}$$

The **line integral of $f(x,y,z)$ along C with respect to x** is

$$\int_C f(x,y,z)\,dx = \int_a^b f(x(t),y(t),z(t))\,x'(t)\,dt\,. \tag{4.35}$$

The **line integral of $f(x,y,z)$ along C with respect to y** is

$$\int_C f(x,y,z)\,dy = \int_a^b f(x(t),y(t),z(t))\,y'(t)\,dt\,. \tag{4.36}$$

The **line integral of $f(x,y,z)$ along C with respect to z** is

$$\int_C f(x,y,z)\,dz = \int_a^b f(x(t),y(t),z(t))\,z'(t)\,dt\,. \tag{4.37}$$

Similar to the two-variable case, if $f(x,y,z) \ge 0$ then the line integral $\int_C f(x,y,z)\,ds$ can be thought of as the total area of the "picket fence" of height $f(x,y,z)$ at each point along the curve C in $\mathbb{R}^3$.

Vector fields in $\mathbb{R}^3$ are defined in a similar fashion to those in $\mathbb{R}^2$, which allows us to define the line integral of a vector field along a curve in $\mathbb{R}^3$.

Definition 4.6. For a vector field $\mathbf{f}(x,y,z) = P(x,y,z)\mathbf{i}+Q(x,y,z)\mathbf{j}+R(x,y,z)\mathbf{k}$ and a curve C in $\mathbb{R}^3$ with a smooth parametrization $x = x(t)$, $y = y(t)$, $z = z(t)$, $a \le t \le b$, the **line integral of f along** C is

$$\int_C \mathbf{f}\cdot d\mathbf{r} = \int_C P(x,y,z)\,dx + \int_C Q(x,y,z)\,dy + \int_C R(x,y,z)\,dz \tag{4.38}$$

$$= \int_a^b \mathbf{f}(x(t),y(t),z(t))\cdot\mathbf{r}'(t)\,dt\,, \tag{4.39}$$

where $\mathbf{r}(t) = x(t)\mathbf{i}+y(t)\mathbf{j}+z(t)\mathbf{k}$ is the position vector for points on C.

Similar to the two-variable case, if $\mathbf{f}(x,y,z)$ represents the force applied to an object at a point (x,y,z) then the line integral $\int_C \mathbf{f}\cdot d\mathbf{r}$ represents the work done by that force in moving the object along the curve C in $\mathbb{R}^3$.

Some of the most important results we will need for line integrals in $\mathbb{R}^3$ are stated below without proof (the proofs are similar to their two-variable equivalents).

Theorem 4.10. For a vector field $\mathbf{f}(x,y,z) = P(x,y,z)\,\mathbf{i} + Q(x,y,z)\,\mathbf{j} + R(x,y,z)\,\mathbf{k}$ and a curve C with a smooth parametrization $x = x(t)$, $y = y(t)$, $z = z(t)$, $a \le t \le b$ and position vector $\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j} + z(t)\,\mathbf{k}$,

$$\int_C \mathbf{f}\cdot d\mathbf{r} \;=\; \int_C \mathbf{f}\cdot\mathbf{T}\,ds\,, \tag{4.40}$$

where $\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{\|\mathbf{r}'(t)\|}$ is the unit tangent vector to C at $(x(t),y(t),z(t))$.

Theorem 4.11. **(Chain Rule)** If $w = f(x,y,z)$ is a continuously differentiable function of x, y, and z, and $x = x(t)$, $y = y(t)$ and $z = z(t)$ are differentiable functions of t, then w is a differentiable function of t, and

$$\frac{dw}{dt} \;=\; \frac{\partial w}{\partial x}\frac{dx}{dt} \;+\; \frac{\partial w}{\partial y}\frac{dy}{dt} \;+\; \frac{\partial w}{\partial z}\frac{dz}{dt}\,. \tag{4.41}$$

Also, if $x = x(t_1,t_2)$, $y = y(t_1,t_2)$ and $z = z(t_1,t_2)$ are continuously differentiable function of (t_1,t_2), then[6]

$$\frac{\partial w}{\partial t_1} \;=\; \frac{\partial w}{\partial x}\frac{\partial x}{\partial t_1} \;+\; \frac{\partial w}{\partial y}\frac{\partial y}{\partial t_1} \;+\; \frac{\partial w}{\partial z}\frac{\partial z}{\partial t_1} \tag{4.42}$$

and

$$\frac{\partial w}{\partial t_2} \;=\; \frac{\partial w}{\partial x}\frac{\partial x}{\partial t_2} \;+\; \frac{\partial w}{\partial y}\frac{\partial y}{\partial t_2} \;+\; \frac{\partial w}{\partial z}\frac{\partial z}{\partial t_2}\,. \tag{4.43}$$

Theorem 4.12. Let $\mathbf{f}(x,y,z) = P(x,y,z)\,\mathbf{i} + Q(x,y,z)\,\mathbf{j} + R(x,y,z)\,\mathbf{k}$ be a vector field in some solid S, with P, Q and R continuously differentiable functions on S. Let C be a smooth curve in S parametrized by $x = x(t)$, $y = y(t)$, $z = z(t)$, $a \le t \le b$. Suppose that there is a real-valued function $F(x,y,z)$ such that $\nabla F = \mathbf{f}$ on S. Then

$$\int_C \mathbf{f}\cdot d\mathbf{r} \;=\; F(B) \;-\; F(A)\,, \tag{4.44}$$

where $A = (x(a),y(a),z(a))$ and $B = (x(b),y(b),z(b))$ are the endpoints of C.

Corollary 4.13. If a vector field $\mathbf{f}$ has a potential in a solid S, then $\oint_C \mathbf{f}\cdot d\mathbf{r} = 0$ for any closed curve C in S (i.e. $\oint_C \nabla F\cdot d\mathbf{r} = 0$ for any real-valued function $F(x,y,z)$).

[6]See TAYLOR and MANN, § 6.5 for a proof.

Example 4.12. Let $f(x, y, z) = z$ and let C be the curve in $\mathbb{R}^3$ parametrized by

$$x = t\sin t\,, \quad y = t\cos t\,, \quad z = t\,, \quad 0 \le t \le 8\pi\,.$$

Evaluate $\int_C f(x, y, z)\,ds$. (Note: C is called a *conical helix*. See Figure 4.5.1).

Solution: Since $x'(t) = \sin t + t\cos t$, $y'(t) = \cos t - t\sin t$, and $z'(t) = 1$, we have

$$\begin{aligned}
x'(t)^2 + y'(t)^2 + z'(t)^2 &= (\sin^2 t + 2t\sin t\cos t + t^2\cos^2 t) + (\cos^2 t - 2t\sin t\cos t + t^2\sin^2 t) + 1\\
&= t^2(\sin^2 t + \cos^2 t) + \sin^2 t + \cos^2 t + 1\\
&= t^2 + 2\,,
\end{aligned}$$

so since $f(x(t), y(t), z(t)) = z(t) = t$ along the curve C, then

$$\begin{aligned}
\int_C f(x, y, z)\,ds &= \int_0^{8\pi} f(x(t), y(t), z(t))\sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\,dt\\
&= \int_0^{8\pi} t\sqrt{t^2 + 2}\,dt\\
&= \left(\frac{1}{3}(t^2 + 2)^{3/2}\right)\Bigg|_0^{8\pi} = \frac{1}{3}\left((64\pi^2 + 2)^{3/2} - 2\sqrt{2}\right).
\end{aligned}$$

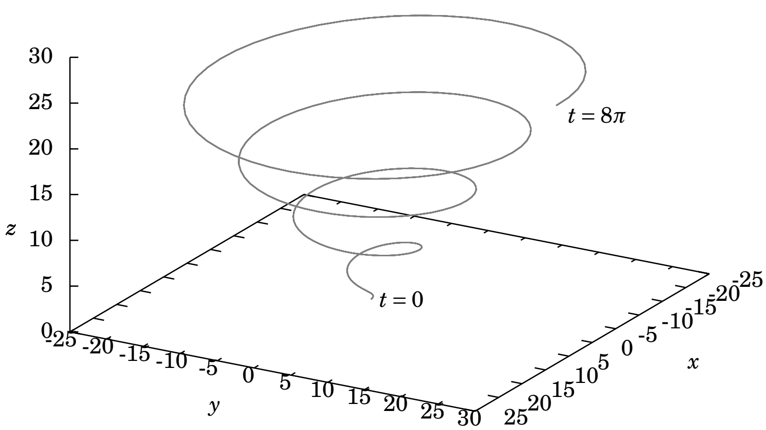


Figure 4.5.1 Conical helix C

Example 4.13. Let $\mathbf{f}(x, y, z) = x\,\mathbf{i} + y\,\mathbf{j} + 2z\,\mathbf{k}$ be a vector field in $\mathbb{R}^3$. Using the same curve C from Example 4.12, evaluate $\int_C \mathbf{f} \cdot d\mathbf{r}$.

Solution: It is easy to see that $F(x, y, z) = \frac{x^2}{2} + \frac{y^2}{2} + z^2$ is a potential for $\mathbf{f}(x, y, z)$ (i.e. $\nabla F = \mathbf{f}$).

So by Theorem 4.12 we know that

$$\begin{aligned}\int_C \mathbf{f}\cdot d\mathbf{r} &= F(B) - F(A)\ \text{, where } A = (x(0), y(0), z(0)) \text{ and } B = (x(8\pi), y(8\pi), z(8\pi)), \text{ so}\\ &= F(8\pi\sin 8\pi, 8\pi\cos 8\pi, 8\pi) - F(0\sin 0, 0\cos 0, 0)\\ &= F(0, 8\pi, 8\pi) - F(0,0,0)\\ &= 0 + \frac{(8\pi)^2}{2} + (8\pi)^2 - (0+0+0) = 96\pi^2\ .\end{aligned}$$

We will now discuss a generalization of Green's Theorem in $\mathbb{R}^2$ to *orientable* surfaces in $\mathbb{R}^3$, called *Stokes' Theorem*. A surface Σ in $\mathbb{R}^3$ is **orientable** if there is a continuous vector field $\mathbf{N}$ in $\mathbb{R}^3$ such that $\mathbf{N}$ is nonzero and normal to Σ (i.e. perpendicular to the tangent plane) at each point of Σ. We say that such an $\mathbf{N}$ is a *normal vector field*.

For example, the unit sphere $x^2+y^2+z^2=1$ is orientable, since the continuous vector field $\mathbf{N}(x,y,z) = x\,\mathbf{i}+y\,\mathbf{j}+z\,\mathbf{k}$ is nonzero and normal to the sphere at each point. In fact, $-\mathbf{N}(x,y,z)$ is another normal vector field (see Figure 4.5.2). We see in this case that $\mathbf{N}(x,y,z)$ is what we have called an outward normal vector, and $-\mathbf{N}(x,y,z)$ is an *inward* normal vector. These "outward" and "inward" normal vector fields on the sphere correspond to an "outer" and "inner" side, respectively, of the sphere. That is, we say that the sphere is a *two-sided* surface. Roughly, "two-sided" means "orientable". Other examples of two-sided, and hence orientable, surfaces are cylinders, paraboloids, ellipsoids, and planes.

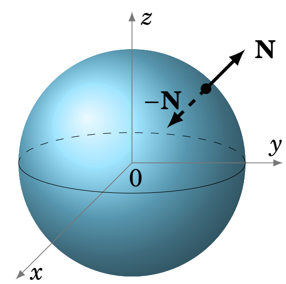
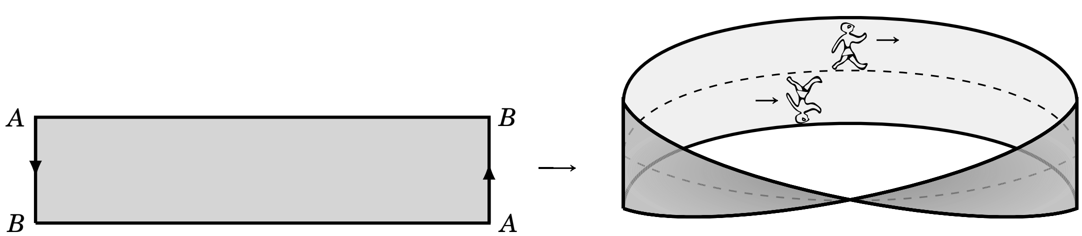


Figure 4.5.2

You may be wondering what kind of surface would *not* have two sides. An example is the **Möbius strip**, which is constructed by taking a thin rectangle and connecting its ends at the opposite corners, resulting in a "twisted" strip (see Figure 4.5.3).



(a) Connect A to A and B to B along the ends

(b) Not orientable

Figure 4.5.3 Möbius strip

If you imagine walking along a line down the center of the Möbius strip, as in Figure 4.5.3(b), then you arrive back at the same place from which you started but upside down! That is, your *orientation* changed even though your motion was continuous along that center

line. Informally, thinking of your vertical direction as a normal vector field along the strip, there is a discontinuity at your starting point (and, in fact, at every point) since your vertical direction takes two different values there. The Möbius strip has only *one side*, and hence is nonorientable.[7]

For an orientable surface Σ which has a boundary curve C, pick a unit normal vector $\mathbf{n}$ such that if you walked along C with your head pointing in the direction of $\mathbf{n}$, then the surface would be on your left. We say in this situation that $\mathbf{n}$ is a *positive unit normal vector* and that C is traversed ***n**-positively*. We can now state Stokes' Theorem:

Theorem 4.14. **(Stokes' Theorem)** Let Σ be an orientable surface in $\mathbb{R}^3$ whose boundary is a simple closed curve C, and let $\mathbf{f}(x,y,z) = P(x,y,z)\mathbf{i} + Q(x,y,z)\mathbf{j} + R(x,y,z)\mathbf{k}$ be a smooth vector field defined on some subset of $\mathbb{R}^3$ that contains Σ. Then

$$\oint_C \mathbf{f}\cdot d\mathbf{r} = \iint_\Sigma (\text{curl}\ \mathbf{f})\cdot\mathbf{n}\, d\sigma\ , \tag{4.45}$$

where

$$\text{curl}\ \mathbf{f} = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)\mathbf{i} + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)\mathbf{j} + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathbf{k}\ , \tag{4.46}$$

$\mathbf{n}$ is a positive unit normal vector over Σ, and C is traversed $\mathbf{n}$-positively.

Proof: As the general case is beyond the scope of this text, we will prove the theorem only for the special case where Σ is the graph of $z = z(x,y)$ for some smooth real-valued function $z(x,y)$, with (x,y) varying over a region D in $\mathbb{R}^2$.

Projecting Σ onto the xy-plane, we see that the closed curve C (the boundary curve of Σ) projects onto a closed curve C_D which is the boundary curve of D (see Figure 4.5.4). Assuming that C has a smooth parametrization, its projection C_D in the xy-plane also has a smooth parametrization, say

$$C_D:\ x = x(t)\ ,\ y = y(t)\ ,\ a \le t \le b\ ,$$

and so C can be parametrized (in $\mathbb{R}^3$) as

$$C:\ x = x(t)\ ,\ y = y(t)\ ,\ z = z(x(t), y(t))\ ,\ a \le t \le b\ ,$$

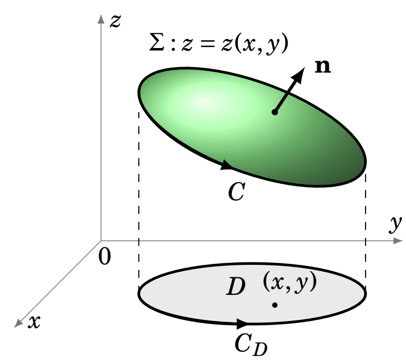


Figure 4.5.4

since the curve C is part of the surface $z = z(x,y)$. Now, by the Chain Rule (Theorem 4.4 in Section 4.2), for $z = z(x(t), y(t))$ as a function of t, we know that

$$z'(t) = \frac{\partial z}{\partial x}x'(t) + \frac{\partial z}{\partial y}y'(t)\ ,$$

[7]For further discussion of orientability, see O'NEILL, § IV.7.

and so

$$\begin{aligned}
\oint_C \mathbf{f}\cdot d\mathbf{r} &= \int_C P(x,y,z)\,dx + Q(x,y,z)\,dy + R(x,y,z)\,dz \\
&= \int_a^b \left(P\,x'(t) + Q\,y'(t) + R\left(\frac{\partial z}{\partial x}\,x'(t) + \frac{\partial z}{\partial y}\,y'(t)\right)\right) dt \\
&= \int_a^b \left(\left(P + R\,\frac{\partial z}{\partial x}\right)x'(t) + \left(Q + R\,\frac{\partial z}{\partial y}\right)y'(t)\right) dt \\
&= \int_{C_D} \tilde{P}(x,y)\,dx + \tilde{Q}(x,y)\,dy\ ,
\end{aligned}$$

where

$$\begin{aligned}
\tilde{P}(x,y) &= P(x,y,z(x,y)) \;+\; R(x,y,z(x,y))\,\frac{\partial z}{\partial x}(x,y)\ ,\text{and} \\
\tilde{Q}(x,y) &= Q(x,y,z(x,y)) \;+\; R(x,y,z(x,y))\,\frac{\partial z}{\partial y}(x,y)
\end{aligned}$$

for (x,y) in D. Thus, by Green's Theorem applied to the region D, we have

$$\oint_C \mathbf{f}\cdot d\mathbf{r} \;=\; \iint\limits_D \left(\frac{\partial \tilde{Q}}{\partial x} - \frac{\partial \tilde{P}}{\partial y}\right) dA\ . \tag{4.47}$$

Thus,

$$\begin{aligned}
\frac{\partial \tilde{Q}}{\partial x} &= \frac{\partial}{\partial x}\left(Q(x,y,z(x,y)) + R(x,y,z(x,y))\,\frac{\partial z}{\partial y}(x,y)\right)\ \text{, so by the Product Rule we get} \\
&= \frac{\partial}{\partial x}\left(Q(x,y,z(x,y))\right) + \left(\frac{\partial}{\partial x}R(x,y,z(x,y))\right)\frac{\partial z}{\partial y}(x,y) + R(x,y,z(x,y))\,\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}(x,y)\right)\ .
\end{aligned}$$

Now, by formula (4.42) in Theorem 4.11, we have

$$\begin{aligned}
\frac{\partial}{\partial x}\left(Q(x,y,z(x,y))\right) &= \frac{\partial Q}{\partial x}\,\frac{\partial x}{\partial x} \;+\; \frac{\partial Q}{\partial y}\,\frac{\partial y}{\partial x} \;+\; \frac{\partial Q}{\partial z}\,\frac{\partial z}{\partial x} \\
&= \frac{\partial Q}{\partial x}\cdot 1 \;+\; \frac{\partial Q}{\partial y}\cdot 0 \;+\; \frac{\partial Q}{\partial z}\,\frac{\partial z}{\partial x} \\
&= \frac{\partial Q}{\partial x} \;+\; \frac{\partial Q}{\partial z}\,\frac{\partial z}{\partial x}\ .
\end{aligned}$$

Similarly,

$$\frac{\partial}{\partial x}\left(R(x,y,z(x,y))\right) \;=\; \frac{\partial R}{\partial x} \;+\; \frac{\partial R}{\partial z}\,\frac{\partial z}{\partial x}\ .$$

Thus,

$$\frac{\partial \tilde{Q}}{\partial x} = \frac{\partial Q}{\partial x} + \frac{\partial Q}{\partial z}\frac{\partial z}{\partial x} + \left(\frac{\partial R}{\partial x} + \frac{\partial R}{\partial z}\frac{\partial z}{\partial x}\right)\frac{\partial z}{\partial y} + R(x,y,z(x,y))\frac{\partial^2 z}{\partial x\,\partial y}$$
$$= \frac{\partial Q}{\partial x} + \frac{\partial Q}{\partial z}\frac{\partial z}{\partial x} + \frac{\partial R}{\partial x}\frac{\partial z}{\partial y} + \frac{\partial R}{\partial z}\frac{\partial z}{\partial x}\frac{\partial z}{\partial y} + R\frac{\partial^2 z}{\partial x\,\partial y}\,.$$

In a similar fashion, we can calculate

$$\frac{\partial \tilde{P}}{\partial y} = \frac{\partial P}{\partial y} + \frac{\partial P}{\partial z}\frac{\partial z}{\partial y} + \frac{\partial R}{\partial y}\frac{\partial z}{\partial x} + \frac{\partial R}{\partial z}\frac{\partial z}{\partial y}\frac{\partial z}{\partial x} + R\frac{\partial^2 z}{\partial y\,\partial x}\,.$$

So subtracting gives

$$\frac{\partial \tilde{Q}}{\partial x} - \frac{\partial \tilde{P}}{\partial y} = \left(\frac{\partial Q}{\partial z} - \frac{\partial R}{\partial y}\right)\frac{\partial z}{\partial x} + \left(\frac{\partial R}{\partial x} - \frac{\partial P}{\partial z}\right)\frac{\partial z}{\partial y} + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) \tag{4.48}$$

since $\frac{\partial^2 z}{\partial x\partial y} = \frac{\partial^2 z}{\partial y\partial x}$ by the smoothness of $z = z(x,y)$. Hence, by equation (4.47),

$$\oint_C \mathbf{f}\cdot d\mathbf{r} = \iint_D \left(-\left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)\frac{\partial z}{\partial x} - \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)\frac{\partial z}{\partial y} + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\right) dA \tag{4.49}$$

after factoring out a -1 from the terms in the first two products in equation (4.48).

Now, recall from Section 2.3 (see p.76) that the vector $\mathbf{N} = -\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}$ is normal to the tangent plane to the surface $z = z(x,y)$ at each point of Σ. Thus,

$$\mathbf{n} = \frac{\mathbf{N}}{\|\mathbf{N}\|} = \frac{-\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}}{\sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}}$$

is in fact a positive unit normal vector to Σ (see Figure 4.5.4). Hence, using the parametrization $\mathbf{r}(x,y) = x\,\mathbf{i} + y\,\mathbf{j} + z(x,y)\mathbf{k}$, for (x,y) in D, of the surface Σ, we have $\frac{\partial \mathbf{r}}{\partial x} = \mathbf{i} + \frac{\partial z}{\partial x}\mathbf{k}$ and $\frac{\partial \mathbf{r}}{\partial y} = \mathbf{j} + \frac{\partial z}{\partial y}\mathbf{k}$, and so $\left\|\frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y}\right\| = \sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}$. So we see that using formula (4.46) for curl $\mathbf{f}$, we have

$$\iint_\Sigma (\text{curl }\mathbf{f})\cdot\mathbf{n}\,d\sigma = \iint_D (\text{curl }\mathbf{f})\cdot\mathbf{n}\left\|\frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y}\right\| dA$$
$$= \iint_D \left(\left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)\mathbf{i} + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)\mathbf{j} + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathbf{k}\right)\cdot\left(-\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}\right) dA$$
$$= \iint_D \left(-\left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)\frac{\partial z}{\partial x} - \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)\frac{\partial z}{\partial y} + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\right) dA\,,$$

which, upon comparing to equation (4.49), proves the Theorem. **QED**

Note: The condition in Stokes' Theorem that the surface Σ have a (continuously varying) positive unit normal vector $\mathbf{n}$ and a boundary curve C traversed $\mathbf{n}$-positively can be expressed more precisely as follows: if $\mathbf{r}(t)$ is the position vector for C and $\mathbf{T}(t) = \mathbf{r}'(t)/\|\mathbf{r}'(t)\|$ is the unit tangent vector to C, then the vectors $\mathbf{T}$, $\mathbf{n}$, $\mathbf{T} \times \mathbf{n}$ form a right-handed system.

Also, it should be noted that Stokes' Theorem holds even when the boundary curve C is piecewise smooth.

Example 4.14. Verify Stokes' Theorem for $\mathbf{f}(x, y, z) = z\,\mathbf{i} + x\,\mathbf{j} + y\,\mathbf{k}$ when Σ is the paraboloid $z = x^2 + y^2$ such that $z \le 1$ (see Figure 4.5.5).

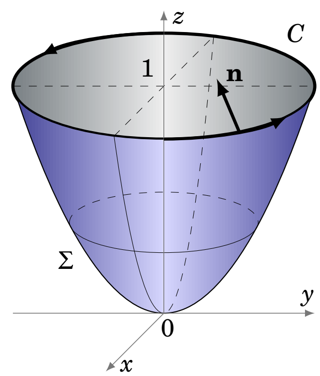


Figure 4.5.5 $z = x^2 + y^2$

Solution: The positive unit normal vector to the surface $z = z(x, y) = x^2 + y^2$ is

$$\mathbf{n} = \frac{-\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}}{\sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}} = \frac{-2x\,\mathbf{i} - 2y\,\mathbf{j} + \mathbf{k}}{\sqrt{1 + 4x^2 + 4y^2}},$$

and $\text{curl}\,\mathbf{f} = (1-0)\,\mathbf{i} + (1-0)\,\mathbf{j} + (1-0)\,\mathbf{k} = \mathbf{i} + \mathbf{j} + \mathbf{k}$, so

$$(\text{curl}\,\mathbf{f}) \cdot \mathbf{n} = (-2x - 2y + 1)/\sqrt{1 + 4x^2 + 4y^2}\,.$$

Since Σ can be parametrized as $\mathbf{r}(x, y) = x\,\mathbf{i} + y\,\mathbf{j} + (x^2 + y^2)\,\mathbf{k}$ for (x, y) in the region $D = \{(x, y) : x^2 + y^2 \le 1\}$, then

$$\begin{aligned}
\iint_{\Sigma} (\text{curl}\,\mathbf{f}) \cdot \mathbf{n}\, d\sigma &= \iint_{D} (\text{curl}\,\mathbf{f}) \cdot \mathbf{n} \left\| \frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y} \right\| dA \\
&= \iint_{D} \frac{-2x - 2y + 1}{\sqrt{1 + 4x^2 + 4y^2}} \sqrt{1 + 4x^2 + 4y^2}\, dA \\
&= \iint_{D} (-2x - 2y + 1)\, dA \text{ , so switching to polar coordinates gives} \\
&= \int_0^{2\pi} \int_0^1 (-2r\cos\theta - 2r\sin\theta + 1)\, r\, dr\, d\theta \\
&= \int_0^{2\pi} \int_0^1 (-2r^2\cos\theta - 2r^2\sin\theta + r)\, dr\, d\theta \\
&= \int_0^{2\pi} \left(-\frac{2r^3}{3}\cos\theta - \frac{2r^3}{3}\sin\theta + \frac{r^2}{2} \Bigg|_{r=0}^{r=1} \right) d\theta \\
&= \int_0^{2\pi} \left(-\frac{2}{3}\cos\theta - \frac{2}{3}\sin\theta + \frac{1}{2} \right) d\theta \\
&= -\frac{2}{3}\sin\theta + \frac{2}{3}\cos\theta + \frac{1}{2}\theta \Bigg|_0^{2\pi} = \pi\,.
\end{aligned}$$

The boundary curve C is the unit circle $x^2 + y^2 = 1$ laying in the plane $z = 1$ (see Figure 4.5.5), which can be parametrized as $x = \cos t$, $y = \sin t$, $z = 1$ for $0 \le t \le 2\pi$. So

$$\begin{aligned}
\oint_C \mathbf{f} \cdot d\mathbf{r} &= \int_0^{2\pi} ((1)(-\sin t) + (\cos t)(\cos t) + (\sin t)(0))\,dt \\
&= \int_0^{2\pi} \left(-\sin t + \frac{1+\cos 2t}{2}\right) dt \quad \left(\text{here we used } \cos^2 t = \frac{1+\cos 2t}{2}\right) \\
&= \cos t + \frac{t}{2} + \frac{\sin 2t}{4} \bigg|_0^{2\pi} = \pi\,.
\end{aligned}$$

So we see that $\oint_C \mathbf{f} \cdot d\mathbf{r} = \iint\limits_{\Sigma} (\text{curl } \mathbf{f}) \cdot \mathbf{n}\, d\sigma$, as predicted by Stokes' Theorem.

The line integral in the preceding example was far simpler to calculate than the surface integral, but this will not always be the case.

Example 4.15. Let Σ be the elliptic paraboloid $z = \frac{x^2}{4} + \frac{y^2}{9}$ for $z \le 1$, and let C be its boundary curve. Calculate $\oint_C \mathbf{f} \cdot d\mathbf{r}$ for $\mathbf{f}(x, y, z) = (9xz + 2y)\mathbf{i} + (2x + y^2)\mathbf{j} + (-2y^2 + 2z)\mathbf{k}$, where C is traversed counterclockwise.

Solution: The surface is similar to the one in Example 4.14, except now the boundary curve C is the ellipse $\frac{x^2}{4} + \frac{y^2}{9} = 1$ laying in the plane $z = 1$. In this case, using Stokes' Theorem is easier than computing the line integral directly. As in Example 4.14, at each point $(x, y, z(x, y))$ on the surface $z = z(x, y) = \frac{x^2}{4} + \frac{y^2}{9}$ the vector

$$\mathbf{n} = \frac{-\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}}{\sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}} = \frac{-\frac{x}{2}\mathbf{i} - \frac{2y}{9}\mathbf{j} + \mathbf{k}}{\sqrt{1 + \frac{x^2}{4} + \frac{4y^2}{9}}}\,,$$

is a positive unit normal vector to Σ. And calculating the curl of $\mathbf{f}$ gives

$$\text{curl } \mathbf{f} = (-4y - 0)\mathbf{i} + (9x - 0)\mathbf{j} + (2 - 2)\mathbf{k} = -4y\,\mathbf{i} + 9x\,\mathbf{j} + 0\,\mathbf{k}\,,$$

so

$$(\text{curl } \mathbf{f}) \cdot \mathbf{n} = \frac{(-4y)(-\frac{x}{2}) + (9x)(-\frac{2y}{9}) + (0)(1)}{\sqrt{1 + \frac{x^2}{4} + \frac{4y^2}{9}}} = \frac{2xy - 2xy + 0}{\sqrt{1 + \frac{x^2}{4} + \frac{4y^2}{9}}} = 0\,,$$

and so by Stokes' Theorem

$$\oint_C \mathbf{f} \cdot d\mathbf{r} = \iint\limits_{\Sigma} (\text{curl } \mathbf{f}) \cdot \mathbf{n}\, d\sigma = \iint\limits_{\Sigma} 0\, d\sigma = 0\,.$$

In physical applications, for a simple closed curve C the line integral $\oint_C \mathbf{f}\cdot d\mathbf{r}$ is often called the **circulation** of $\mathbf{f}$ around C. For example, if $\mathbf{E}$ represents the electrostatic field due to a point charge, then it turns out[8] that curl $\mathbf{E} = \mathbf{0}$, which means that the circulation $\oint_C \mathbf{E}\cdot d\mathbf{r} = 0$ by Stokes' Theorem. Vector fields which have zero curl are often called *irrotational* fields.

In fact, the term curl was created by the 19th century Scottish physicist James Clerk Maxwell in his study of electromagnetism, where it is used extensively. In physics, the curl is interpreted as a measure of *circulation density*. This is best seen by using another definition of curl $\mathbf{f}$ which is equivalent[9] to the definition given by formula (4.46). Namely, for a point (x,y,z) in $\mathbb{R}^3$,

$$\mathbf{n}\cdot(\text{curl }\mathbf{f})(x,y,z) \;=\; \lim_{S\to 0}\frac{1}{S}\oint_C \mathbf{f}\cdot d\mathbf{r}\;, \tag{4.50}$$

where S is the surface area of a surface Σ containing the point (x,y,z) and with a simple closed boundary curve C and positive unit normal vector $\mathbf{n}$ at (x,y,z). In the limit, think of the curve C shrinking to the point (x,y,z), which causes Σ, the surface it bounds, to have smaller and smaller surface area. That ratio of circulation to surface area in the limit is what makes the curl a rough measure of circulation density (i.e. circulation per unit area).

An idea of how the curl of a vector field is related to rotation is shown in Figure 4.5.6. Suppose we have a vector field $\mathbf{f}(x,y,z)$ which is always parallel to the xy-plane at each point (x,y,z) and that the vectors grow larger the further the point (x,y,z) is from the y-axis. For example, $\mathbf{f}(x,y,z) = (1+x^2)\mathbf{j}$. Think of the vector field as representing the flow of water, and imagine dropping two wheels with paddles into that water flow, as in Figure 4.5.6. Since the flow is stronger (i.e. the magnitude of $\mathbf{f}$ is larger) as you move away from the y-axis, then such a wheel would rotate counterclockwise if it were dropped to the right of the y-axis, and it would rotate clockwise if it were dropped to the left of the y-axis. In both cases the curl would be nonzero (curl $\mathbf{f}(x,y,z) = 2x\,\mathbf{k}$ in our example) and would obey the right-hand rule, that is, curl $\mathbf{f}(x,y,z)$ points in the direction of your thumb as you cup your right hand in the direction of the rotation of the wheel. So the curl points outward (in the positive z-direction) if $x > 0$ and points inward (in the negative z-direction) if $x < 0$. Notice that if all the vectors had the same direction *and* the same magnitude, then the wheels would not rotate and hence there would be no curl (which is why such fields are called irrotational, meaning no rotation).

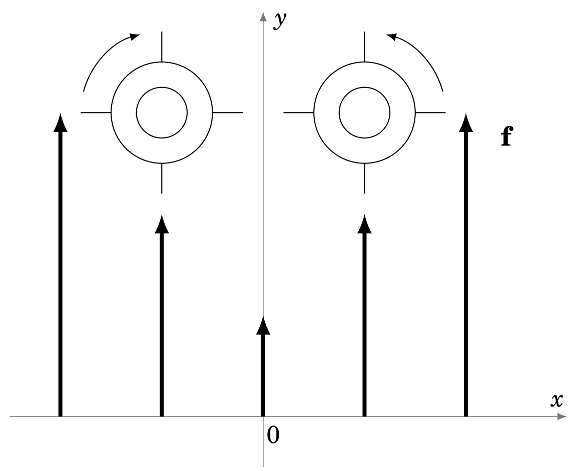


Figure 4.5.6 Curl and rotation

[8]See Ch. 2 in REITZ, MILFORD and CHRISTY.

[9]See SCHEY, p. 78-81, for the derivation.

Finally, by Stokes' Theorem, we know that if C is a simple closed curve in some solid region S in $\mathbb{R}^3$ and if $\mathbf{f}(x,y,z)$ is a smooth vector field such that curl $\mathbf{f} = \mathbf{0}$ in S, then

$$\oint_C \mathbf{f} \cdot d\mathbf{r} \;=\; \iint_\Sigma (\text{curl}\,\mathbf{f}) \cdot \mathbf{n}\,d\sigma \;=\; \iint_\Sigma \mathbf{0} \cdot \mathbf{n}\,d\sigma \;=\; \iint_\Sigma 0\,d\sigma \;=\; 0\,,$$

where Σ is any orientable surface inside S whose boundary is C (such a surface is sometimes called a *capping surface* for C). So similar to the two-variable case, we have a three-dimensional version of a result from Section 4.3, for solid regions in $\mathbb{R}^3$ which are **simply connected** (i.e. regions having no holes):

The following statements are equivalent for a simply connected solid region S in $\mathbb{R}^3$:

(a) $\mathbf{f}(x,y,z) = P(x,y,z)\,\mathbf{i} + Q(x,y,z)\,\mathbf{j} + R(x,y,z)\,\mathbf{k}$ has a smooth potential $F(x,y,z)$ in S

(b) $\displaystyle\int_C \mathbf{f} \cdot d\mathbf{r}$ is independent of the path for any curve C in S

(c) $\displaystyle\oint_C \mathbf{f} \cdot d\mathbf{r} = 0$ for every simple closed curve C in S

(d) $\dfrac{\partial R}{\partial y} = \dfrac{\partial Q}{\partial z}$, $\dfrac{\partial P}{\partial z} = \dfrac{\partial R}{\partial x}$, and $\dfrac{\partial Q}{\partial x} = \dfrac{\partial P}{\partial y}$ in S (i.e. curl $\mathbf{f} = \mathbf{0}$ in S)

Part (d) is also a way of saying that the differential form $P\,dx + Q\,dy + R\,dz$ is exact.

Example 4.16. Determine if the vector field $\mathbf{f}(x,y,z) = xyz\,\mathbf{i} + xz\,\mathbf{j} + xy\,\mathbf{k}$ has a potential in $\mathbb{R}^3$.

Solution: Since $\mathbb{R}^3$ is simply connected, we just need to check whether curl $\mathbf{f} = \mathbf{0}$ throughout $\mathbb{R}^3$, that is,

$$\frac{\partial R}{\partial y} = \frac{\partial Q}{\partial z}\,, \quad \frac{\partial P}{\partial z} = \frac{\partial R}{\partial x}\,, \quad \text{and} \quad \frac{\partial Q}{\partial x} = \frac{\partial P}{\partial y}$$

throughout $\mathbb{R}^3$, where $P(x,y,z) = xyz$, $Q(x,y,z) = xz$, and $R(x,y,z) = xy$. But we see that

$$\frac{\partial P}{\partial z} = xy\,,\; \frac{\partial R}{\partial x} = y \quad \Rightarrow \quad \frac{\partial P}{\partial z} \neq \frac{\partial R}{\partial x} \;\text{ for some } (x,y,z) \text{ in } \mathbb{R}^3.$$

Thus, $\mathbf{f}(x,y,z)$ does not have a potential in $\mathbb{R}^3$.

Exercises

A

For Exercises 1-3, calculate $\int_C f(x,y,z)\,ds$ for the given function $f(x,y,z)$ and curve C.

1. $f(x,y,z) = z$; $\quad C: x = \cos t,\ y = \sin t,\ z = t,\ 0 \le t \le 2\pi$

2. $f(x,y,z) = \dfrac{x}{y} + y + 2yz; \quad C: x = t^2,\ y = t,\ z = 1,\ 1 \le t \le 2$

3. $f(x,y,z) = z^2; \quad C: x = t\sin t,\ y = t\cos t,\ z = \frac{2\sqrt{2}}{3}t^{3/2},\ 0 \le t \le 1$

For Exercises 4-9, calculate $\int_C \mathbf{f} \cdot d\mathbf{r}$ for the given vector field $\mathbf{f}(x,y,z)$ and curve C.

4. $\mathbf{f}(x,y,z) = \mathbf{i} - \mathbf{j} + \mathbf{k}; \quad C: x = 3t,\ y = 2t,\ z = t,\ 0 \le t \le 1$

5. $\mathbf{f}(x,y,z) = y\,\mathbf{i} - x\,\mathbf{j} + z\,\mathbf{k}; \quad C: x = \cos t,\ y = \sin t,\ z = t,\ 0 \le t \le 2\pi$

6. $\mathbf{f}(x,y,z) = x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}; \quad C: x = \cos t,\ y = \sin t,\ z = 2,\ 0 \le t \le 2\pi$

7. $\mathbf{f}(x,y,z) = (y-2z)\,\mathbf{i} + xy\,\mathbf{j} + (2xz+y)\,\mathbf{k}; \quad C: x = t,\ y = 2t,\ z = t^2 - 1,\ 0 \le t \le 1$

8. $\mathbf{f}(x,y,z) = yz\,\mathbf{i} + xz\,\mathbf{j} + xy\,\mathbf{k}; \quad C$: the polygonal path from $(0,0,0)$ to $(1,0,0)$ to $(1,2,0)$

9. $\mathbf{f}(x,y,z) = xy\,\mathbf{i} + (z-x)\,\mathbf{j} + 2yz\,\mathbf{k}; \quad C$: the polygonal path from $(0,0,0)$ to $(1,0,0)$ to $(1,2,0)$ to $(1,2,-2)$

For Exercises 10-13, state whether or not the vector field $\mathbf{f}(x,y,z)$ has a potential in $\mathbb{R}^3$ (you do not need to find the potential itself).

10. $\mathbf{f}(x,y,z) = y\,\mathbf{i} - x\,\mathbf{j} + z\,\mathbf{k}$

11. $\mathbf{f}(x,y,z) = a\,\mathbf{i} + b\,\mathbf{j} + c\,\mathbf{k}$ (a, b, c constant)

12. $\mathbf{f}(x,y,z) = (x+y)\,\mathbf{i} + x\,\mathbf{j} + z^2\,\mathbf{k}$

13. $\mathbf{f}(x,y,z) = xy\,\mathbf{i} - (x - yz^2)\,\mathbf{j} + y^2z\,\mathbf{k}$

B

For Exercises 14-15, verify Stokes' Theorem for the given vector field $\mathbf{f}(x,y,z)$ and surface Σ.

14. $\mathbf{f}(x,y,z) = 2y\,\mathbf{i} - x\,\mathbf{j} + z\,\mathbf{k}; \quad \Sigma: x^2 + y^2 + z^2 = 1,\ z \ge 0$

15. $\mathbf{f}(x,y,z) = xy\,\mathbf{i} + xz\,\mathbf{j} + yz\,\mathbf{k}; \quad \Sigma: z = x^2 + y^2,\ z \le 1$

16. Construct a Möbius strip from a piece of paper, then draw a line down its center (like the dotted line in Figure 4.5.3(b)). Cut the Möbius strip along that center line completely around the strip. How many surfaces does this result in? How would you describe them? Are they orientable?

17. Use Gnuplot (see Appendix C) to plot the Möbius strip parametrized as:

$$\mathbf{r}(u,v) \;=\; \cos u\,(1 + v\cos\tfrac{u}{2})\,\mathbf{i} + \sin u\,(1 + v\cos\tfrac{u}{2})\,\mathbf{j} + v\sin\tfrac{u}{2}\,\mathbf{k}\,,\quad 0 \le u \le 2\pi\,,\quad -\tfrac{1}{2} \le v \le \tfrac{1}{2}$$

C

18. Let Σ be a closed surface and $\mathbf{f}(x,y,z)$ a smooth vector field. Show that $\iint_\Sigma (\text{curl}\ \mathbf{f}) \cdot \mathbf{n}\, d\sigma = 0$. (*Hint: Split Σ in half.*)

19. Show that Green's Theorem is a special case of Stokes' Theorem.

4.6 Gradient, Divergence, Curl and Laplacian

In this final section we will establish some relationships between the gradient, divergence and curl, and we will also introduce a new quantity called the *Laplacian*. We will then show how to write these quantities in cylindrical and spherical coordinates.

For a real-valued function $f(x,y,z)$ on $\mathbb{R}^3$, the gradient $\nabla f(x,y,z)$ is a vector-valued function on $\mathbb{R}^3$, that is, its value at a point (x,y,z) is the vector

$$\nabla f(x,y,z) = \left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y},\frac{\partial f}{\partial z}\right) = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

in $\mathbb{R}^3$, where each of the partial derivatives is evaluated at the point (x,y,z). So in this way, you can think of the *symbol* ∇ as being "applied" to a real-valued function f to produce a vector ∇f.

It turns out that the divergence and curl can also be expressed in terms of the symbol ∇. This is done by thinking of ∇ as a *vector* in $\mathbb{R}^3$, namely

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\,. \tag{4.51}$$

Here, the symbols $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial z}$ are to be thought of as "partial derivative operators" that will get "applied" to a real-valued function, say $f(x,y,z)$, to produce the partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial f}{\partial z}$. For instance, $\frac{\partial}{\partial x}$ "applied" to $f(x,y,z)$ produces $\frac{\partial f}{\partial x}$.

Is ∇ *really* a vector? Strictly speaking, no, since $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial z}$ are not actual numbers. But it helps to *think* of ∇ as a vector, especially with the divergence and curl, as we will soon see. The process of "applying" $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, $\frac{\partial}{\partial z}$ to a real-valued function $f(x,y,z)$ is normally thought of as *multiplying* the quantities:

$$\left(\frac{\partial}{\partial x}\right)(f) = \frac{\partial f}{\partial x}\,, \quad \left(\frac{\partial}{\partial y}\right)(f) = \frac{\partial f}{\partial y}\,, \quad \left(\frac{\partial}{\partial z}\right)(f) = \frac{\partial f}{\partial z}$$

For this reason, ∇ is often referred to as the "del operator", since it "operates" on functions.

For example, it is often convenient to write the divergence div $\mathbf{f}$ as $\nabla \cdot \mathbf{f}$, since for a vector field $\mathbf{f}(x,y,z) = f_1(x,y,z)\mathbf{i} + f_2(x,y,z)\mathbf{j} + f_3(x,y,z)\mathbf{k}$, the dot product of $\mathbf{f}$ with ∇ (thought of as a vector) makes sense:

$$\begin{aligned}
\nabla \cdot \mathbf{f} &= \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\right)\cdot (f_1(x,y,z)\mathbf{i} + f_2(x,y,z)\mathbf{j} + f_3(x,y,z)\mathbf{k})\\
&= \left(\frac{\partial}{\partial x}\right)(f_1) + \left(\frac{\partial}{\partial y}\right)(f_2) + \left(\frac{\partial}{\partial z}\right)(f_3)\\
&= \frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y} + \frac{\partial f_3}{\partial z}\\
&= \operatorname{div}\mathbf{f}
\end{aligned}$$

We can also write curl $\mathbf{f}$ in terms of ∇, namely as $\nabla\times\mathbf{f}$, since for a vector field $\mathbf{f}(x,y,z) = P(x,y,z)\mathbf{i}+Q(x,y,z)\mathbf{j}+R(x,y,z)\mathbf{k}$, we have:

$$\begin{aligned}
\nabla\times\mathbf{f} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ P(x,y,z) & Q(x,y,z) & R(x,y,z) \end{vmatrix} \\
&= \left(\frac{\partial R}{\partial y}-\frac{\partial Q}{\partial z}\right)\mathbf{i} - \left(\frac{\partial R}{\partial x}-\frac{\partial P}{\partial z}\right)\mathbf{j} + \left(\frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y}\right)\mathbf{k} \\
&= \left(\frac{\partial R}{\partial y}-\frac{\partial Q}{\partial z}\right)\mathbf{i} + \left(\frac{\partial P}{\partial z}-\frac{\partial R}{\partial x}\right)\mathbf{j} + \left(\frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y}\right)\mathbf{k} \\
&= \text{curl }\mathbf{f}
\end{aligned}$$

For a real-valued function $f(x,y,z)$, the gradient $\nabla f(x,y,z) = \frac{\partial f}{\partial x}\mathbf{i}+\frac{\partial f}{\partial y}\mathbf{j}+\frac{\partial f}{\partial z}\mathbf{k}$ is a vector field, so we can take its divergence:

$$\begin{aligned}
\text{div }\nabla f &= \nabla\cdot\nabla f \\
&= \left(\frac{\partial}{\partial x}\mathbf{i}+\frac{\partial}{\partial y}\mathbf{j}+\frac{\partial}{\partial z}\mathbf{k}\right)\cdot\left(\frac{\partial f}{\partial x}\mathbf{i}+\frac{\partial f}{\partial y}\mathbf{j}+\frac{\partial f}{\partial z}\mathbf{k}\right) \\
&= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)+\frac{\partial}{\partial z}\left(\frac{\partial f}{\partial z}\right) \\
&= \frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2}
\end{aligned}$$

Note that this is a real-valued function, to which we will give a special name:

Definition 4.7. For a real-valued function $f(x,y,z)$, the **Laplacian** of f, denoted by Δf, is given by

$$\Delta f(x,y,z) = \nabla\cdot\nabla f = \frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2}\,. \tag{4.52}$$

Often the notation $\nabla^2 f$ is used for the Laplacian instead of Δf, using the convention $\nabla^2 = \nabla\cdot\nabla$.

Example 4.17. Let $\mathbf{r}(x,y,z) = x\,\mathbf{i}+y\,\mathbf{j}+z\,\mathbf{k}$ be the position vector field on $\mathbb{R}^3$. Then $\|\mathbf{r}(x,y,z)\|^2 = \mathbf{r}\cdot\mathbf{r} = x^2+y^2+z^2$ is a real-valued function. Find

(a) the gradient of $\|\mathbf{r}\|^2$

(b) the divergence of $\mathbf{r}$

(c) the curl of $\mathbf{r}$

(d) the Laplacian of $\|\mathbf{r}\|^2$

Solution: (a) $\nabla\|\mathbf{r}\|^2 = 2x\,\mathbf{i} + 2y\,\mathbf{j} + 2z\,\mathbf{k} = 2\mathbf{r}$

(b) $\nabla\cdot\mathbf{r} = \frac{\partial}{\partial x}(x) + \frac{\partial}{\partial y}(y) + \frac{\partial}{\partial z}(z) = 1+1+1 = 3$

(c)

$$\nabla\times\mathbf{r} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ x & y & z \end{vmatrix} = (0-0)\mathbf{i} - (0-0)\mathbf{j} + (0-0)\mathbf{k} = \mathbf{0}$$

(d) $\Delta\|\mathbf{r}\|^2 = \frac{\partial^2}{\partial x^2}(x^2+y^2+z^2) + \frac{\partial^2}{\partial y^2}(x^2+y^2+z^2) + \frac{\partial^2}{\partial z^2}(x^2+y^2+z^2) = 2+2+2 = 6$

Note that we could have calculated $\Delta\|\mathbf{r}\|^2$ another way, using the ∇ notation along with parts (a) and (b):

$$\Delta\|\mathbf{r}\|^2 = \nabla\cdot\nabla\|\mathbf{r}\|^2 = \nabla\cdot 2\mathbf{r} = 2\nabla\cdot\mathbf{r} = 2(3) = 6$$

Notice that in Example 4.17 if we take the curl of the gradient of $\|\mathbf{r}\|^2$ we get

$$\nabla\times(\nabla\|\mathbf{r}\|^2) = \nabla\times 2\mathbf{r} = 2\nabla\times\mathbf{r} = 2\mathbf{0} = \mathbf{0}\,.$$

The following theorem shows that this will be the case in general:

Theorem 4.15. For any smooth real-valued function $f(x,y,z)$, $\nabla\times(\nabla f) = \mathbf{0}$.

Proof: We see by the smoothness of f that

$$\begin{aligned} \nabla\times(\nabla f) &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} & \dfrac{\partial f}{\partial z} \end{vmatrix} \\ &= \left(\frac{\partial^2 f}{\partial y\,\partial z} - \frac{\partial^2 f}{\partial z\,\partial y}\right)\mathbf{i} - \left(\frac{\partial^2 f}{\partial x\,\partial z} - \frac{\partial^2 f}{\partial z\,\partial x}\right)\mathbf{j} + \left(\frac{\partial^2 f}{\partial x\,\partial y} - \frac{\partial^2 f}{\partial y\,\partial x}\right)\mathbf{k} = \mathbf{0}\,, \end{aligned}$$

since the mixed partial derivatives in each component are equal. **QED**

Corollary 4.16. If a vector field $\mathbf{f}(x,y,z)$ has a potential, then curl $\mathbf{f} = \mathbf{0}$.

Another way of stating Theorem 4.15 is that gradients are irrotational. Also, notice that in Example 4.17 if we take the divergence of the curl of $\mathbf{r}$ we trivially get

$$\nabla\cdot(\nabla\times\mathbf{r}) = \nabla\cdot\mathbf{0} = 0\,.$$

The following theorem shows that this will be the case in general:

Theorem 4.17. For any smooth vector field $\mathbf{f}(x,y,z)$, $\nabla\cdot(\nabla\times\mathbf{f}) = 0$.

The proof is straightforward and left as an exercise for the reader.

Corollary 4.18. The flux of the curl of a smooth vector field $\mathbf{f}(x,y,z)$ through any closed surface is zero.

Proof: Let Σ be a closed surface which bounds a solid S. The flux of $\nabla \times \mathbf{f}$ through Σ is

$$\begin{aligned}
\iint_{\Sigma} (\nabla \times \mathbf{f}) \cdot d\boldsymbol{\sigma} &= \iiint_{S} \nabla \cdot (\nabla \times \mathbf{f})\, dV \quad \text{(by the Divergence Theorem)} \\
&= \iiint_{S} 0\, dV \quad \text{(by Theorem 4.17)} \\
&= 0\,.
\end{aligned}$$

QED

There is another method for proving Theorem 4.15 which can be useful, and is often used in physics. Namely, if the surface integral $\iint_{\Sigma} f(x,y,z)\,d\sigma = 0$ for *all* surfaces Σ in some solid region (usually all of $\mathbb{R}^3$), then we must have $f(x,y,z) = 0$ throughout that region. The proof is not trivial, and physicists do not usually bother to prove it. But the result is true, and can also be applied to double and triple integrals.

For instance, to prove Theorem 4.15, assume that $f(x,y,z)$ is a smooth real-valued function on $\mathbb{R}^3$. Let C be a simple closed curve in $\mathbb{R}^3$ and let Σ be any capping surface for C (i.e. Σ is orientable and its boundary is C). Since ∇f is a vector field, then

$$\begin{aligned}
\iint_{\Sigma} (\nabla \times (\nabla f)) \cdot \mathbf{n}\, d\sigma &= \oint_{C} \nabla f \cdot d\mathbf{r} \quad \text{by Stokes' Theorem, so} \\
&= 0 \quad \text{by Corollary 4.13.}
\end{aligned}$$

Since the choice of Σ was arbitrary, then we must have $(\nabla \times (\nabla f)) \cdot \mathbf{n} = 0$ throughout $\mathbb{R}^3$, where $\mathbf{n}$ is any unit vector. Using $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ in place of $\mathbf{n}$, we see that we must have $\nabla \times (\nabla f) = \mathbf{0}$ in $\mathbb{R}^3$, which completes the proof.

Example 4.18. A system of electric charges has a *charge density* $\rho(x,y,z)$ and produces an electrostatic field $\mathbf{E}(x,y,z)$ at points (x,y,z) in space. *Gauss' Law* states that

$$\iint_{\Sigma} \mathbf{E} \cdot d\boldsymbol{\sigma} = 4\pi \iiint_{S} \rho\, dV$$

for any closed surface Σ which encloses the charges, with S being the solid region enclosed by Σ. Show that $\nabla \cdot \mathbf{E} = 4\pi\rho$. This is one of *Maxwell's Equations*.[10]

[10] In Gaussian (or CGS) units.

Solution: By the Divergence Theorem, we have

$$\begin{aligned}
\iiint_S \nabla\cdot\mathbf{E}\,dV &= \iint_\Sigma \mathbf{E}\cdot d\boldsymbol{\sigma}\\
&= 4\pi\iiint_S \rho\,dV \quad \text{by Gauss' Law, so combining the integrals gives}\\
\iiint_S (\nabla\cdot\mathbf{E}-4\pi\rho)\,dV &= 0 \quad \text{, so}\\
\nabla\cdot\mathbf{E}-4\pi\rho &= 0 \quad \text{since } \Sigma \text{ and hence } S \text{ was arbitrary, so}\\
\nabla\cdot\mathbf{E} &= 4\pi\rho\;.
\end{aligned}$$

Often (especially in physics) it is convenient to use other coordinate systems when dealing with quantities such as the gradient, divergence, curl and Laplacian. We will present the formulas for these in cylindrical and spherical coordinates.

Recall from Section 1.7 that a point (x, y, z) can be represented in cylindrical coordinates (r, θ, z), where $x = r\cos\theta$, $y = r\sin\theta$, $z = z$. At each point (r, θ, z), let $\mathbf{e}_r$, $\mathbf{e}_\theta$, $\mathbf{e}_z$ be unit vectors in the direction of increasing r, θ, z, respectively (see Figure 4.6.1). Then $\mathbf{e}_r$, $\mathbf{e}_\theta$, $\mathbf{e}_z$ form an orthonormal set of vectors. Note, by the right-hand rule, that $\mathbf{e}_z \times \mathbf{e}_r = \mathbf{e}_\theta$.

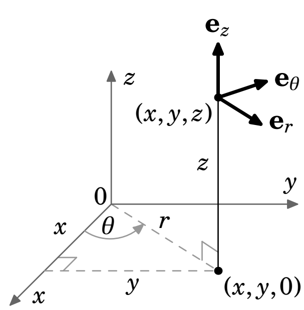


Figure 4.6.1
Orthonormal vectors $\mathbf{e}_r$, $\mathbf{e}_\theta$, $\mathbf{e}_z$ in cylindrical coordinates

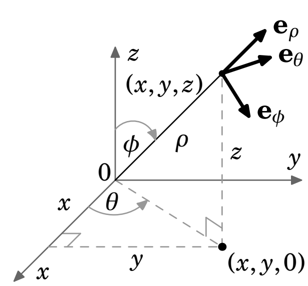


Figure 4.6.2
Orthonormal vectors $\mathbf{e}_\rho$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$ in spherical coordinates

Similarly, a point (x, y, z) can be represented in spherical coordinates (ρ, θ, ϕ), where $x = \rho\sin\phi\cos\theta$, $y = \rho\sin\phi\sin\theta$, $z = \rho\cos\phi$. At each point (ρ, θ, ϕ), let $\mathbf{e}_\rho$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$ be unit vectors in the direction of increasing ρ, θ, ϕ, respectively (see Figure 4.6.2). Then the vectors $\mathbf{e}_\rho$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$ are orthonormal. By the right-hand rule, we see that $\mathbf{e}_\theta \times \mathbf{e}_\rho = \mathbf{e}_\phi$.

We can now summarize the expressions for the gradient, divergence, curl and Laplacian in Cartesian, cylindrical and spherical coordinates in the following tables:

Cartesian (x,y,z): Scalar function F; Vector field $\mathbf{f} = f_1\,\mathbf{i} + f_2\,\mathbf{j} + f_3\,\mathbf{k}$

$$\begin{aligned}
\text{gradient}: \quad \nabla F &= \frac{\partial F}{\partial x}\,\mathbf{i} + \frac{\partial F}{\partial y}\,\mathbf{j} + \frac{\partial F}{\partial z}\,\mathbf{k} \\
\text{divergence}: \quad \nabla \cdot \mathbf{f} &= \frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y} + \frac{\partial f_3}{\partial z} \\
\text{curl}: \nabla \times \mathbf{f} &= \left(\frac{\partial f_3}{\partial y} - \frac{\partial f_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial f_1}{\partial z} - \frac{\partial f_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y}\right)\mathbf{k} \\
\text{Laplacian}: \quad \Delta F &= \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} + \frac{\partial^2 F}{\partial z^2}
\end{aligned}$$

Cylindrical (r,θ,z): Scalar function F; Vector field $\mathbf{f} = f_r\,\mathbf{e}_r + f_\theta\,\mathbf{e}_\theta + f_z\,\mathbf{e}_z$

$$\begin{aligned}
\text{gradient}: \quad \nabla F &= \frac{\partial F}{\partial r}\,\mathbf{e}_r + \frac{1}{r}\frac{\partial F}{\partial \theta}\,\mathbf{e}_\theta + \frac{\partial F}{\partial z}\,\mathbf{e}_z \\
\text{divergence}: \quad \nabla \cdot \mathbf{f} &= \frac{1}{r}\frac{\partial}{\partial r}(r f_r) + \frac{1}{r}\frac{\partial f_\theta}{\partial \theta} + \frac{\partial f_z}{\partial z} \\
\text{curl}: \nabla \times \mathbf{f} &= \left(\frac{1}{r}\frac{\partial f_z}{\partial \theta} - \frac{\partial f_\theta}{\partial z}\right)\mathbf{e}_r + \left(\frac{\partial f_r}{\partial z} - \frac{\partial f_z}{\partial r}\right)\mathbf{e}_\theta + \frac{1}{r}\left(\frac{\partial}{\partial r}(r f_\theta) - \frac{\partial f_r}{\partial \theta}\right)\mathbf{e}_z \\
\text{Laplacian}: \quad \Delta F &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial F}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 F}{\partial \theta^2} + \frac{\partial^2 F}{\partial z^2}
\end{aligned}$$

Spherical (ρ,θ,ϕ): Scalar function F; Vector field $\mathbf{f} = f_\rho\,\mathbf{e}_\rho + f_\theta\,\mathbf{e}_\theta + f_\phi\,\mathbf{e}_\phi$

$$\begin{aligned}
\text{gradient}: \quad \nabla F &= \frac{\partial F}{\partial \rho}\,\mathbf{e}_\rho + \frac{1}{\rho\sin\phi}\frac{\partial F}{\partial \theta}\,\mathbf{e}_\theta + \frac{1}{\rho}\frac{\partial F}{\partial \phi}\,\mathbf{e}_\phi \\
\text{divergence}: \quad \nabla \cdot \mathbf{f} &= \frac{1}{\rho^2}\frac{\partial}{\partial \rho}(\rho^2 f_\rho) + \frac{1}{\rho\sin\phi}\frac{\partial f_\theta}{\partial \theta} + \frac{1}{\rho\sin\phi}\frac{\partial}{\partial \phi}(\sin\phi\, f_\theta) \\
\text{curl}: \nabla \times \mathbf{f} &= \frac{1}{\rho\sin\phi}\left(\frac{\partial}{\partial \phi}(\sin\phi\, f_\theta) - \frac{\partial f_\phi}{\partial \theta}\right)\mathbf{e}_\rho + \frac{1}{\rho}\left(\frac{\partial}{\partial \rho}(\rho f_\phi) - \frac{\partial f_\rho}{\partial \phi}\right)\mathbf{e}_\theta \\
&\quad + \left(\frac{1}{\rho\sin\phi}\frac{\partial f_\rho}{\partial \theta} - \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho f_\theta)\right)\mathbf{e}_\phi \\
\text{Laplacian}: \quad \Delta F &= \frac{1}{\rho^2}\frac{\partial}{\partial \rho}\left(\rho^2\frac{\partial F}{\partial \rho}\right) + \frac{1}{\rho^2\sin^2\phi}\frac{\partial^2 F}{\partial \theta^2} + \frac{1}{\rho^2\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi\,\frac{\partial F}{\partial \phi}\right)
\end{aligned}$$

The derivation of the above formulas for cylindrical and spherical coordinates is straightforward but extremely tedious. The basic idea is to take the Cartesian equivalent of the quantity in question and to substitute into that formula using the appropriate coordinate transformation. As an example, we will derive the formula for the gradient in spherical coordinates.

Goal: Show that the gradient of a real-valued function $F(\rho,\theta,\phi)$ in spherical coordinates is:

$$\nabla F \;=\; \frac{\partial F}{\partial \rho}\mathbf{e}_\rho \;+\; \frac{1}{\rho\sin\phi}\,\frac{\partial F}{\partial \theta}\mathbf{e}_\theta \;+\; \frac{1}{\rho}\,\frac{\partial F}{\partial \phi}\mathbf{e}_\phi$$

Idea: In the Cartesian gradient formula $\nabla F(x,y,z) = \frac{\partial F}{\partial x}\mathbf{i} + \frac{\partial F}{\partial y}\mathbf{j} + \frac{\partial F}{\partial z}\mathbf{k}$, put the Cartesian basis vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ in terms of the spherical coordinate basis vectors $\mathbf{e}_\rho$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$ and functions of ρ, θ and ϕ. Then put the partial derivatives $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, $\frac{\partial F}{\partial z}$ in terms of $\frac{\partial F}{\partial \rho}$, $\frac{\partial F}{\partial \theta}$, $\frac{\partial F}{\partial \phi}$ and functions of ρ, θ and ϕ.

Step 1: Get formulas for $\mathbf{e}_\rho$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$ in terms of $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$.

We can see from Figure 4.6.2 that the unit vector $\mathbf{e}_\rho$ in the ρ direction at a general point (ρ,θ,ϕ) is $\mathbf{e}_\rho = \frac{\mathbf{r}}{\|\mathbf{r}\|}$, where $\mathbf{r} = x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}$ is the position vector of the point in Cartesian coordinates. Thus,

$$\mathbf{e}_\rho \;=\; \frac{\mathbf{r}}{\|\mathbf{r}\|} \;=\; \frac{x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}}{\sqrt{x^2+y^2+z^2}}\;,$$

so using $x = \rho\sin\phi\cos\theta$, $y = \rho\sin\phi\sin\theta$, $z = \rho\cos\phi$, and $\rho = \sqrt{x^2+y^2+z^2}$, we get:

$$\boxed{\mathbf{e}_\rho \;=\; \sin\phi\cos\theta\,\mathbf{i} \;+\; \sin\phi\sin\theta\,\mathbf{j} \;+\; \cos\phi\,\mathbf{k}}$$

Now, since the angle θ is measured in the xy-plane, then the unit vector $\mathbf{e}_\theta$ in the θ direction must be parallel to the xy-plane. That is, $\mathbf{e}_\theta$ is of the form $a\,\mathbf{i} + b\,\mathbf{j} + 0\,\mathbf{k}$. To figure out what a and b are, note that since $\mathbf{e}_\theta \perp \mathbf{e}_\rho$, then in particular $\mathbf{e}_\theta \perp \mathbf{e}_\rho$ when $\mathbf{e}_\rho$ is in the xy-plane. That occurs when the angle ϕ is $\pi/2$. Putting $\phi = \pi/2$ into the formula for $\mathbf{e}_\rho$ gives $\mathbf{e}_\rho = \cos\theta\,\mathbf{i} + \sin\theta\,\mathbf{j} + 0\,\mathbf{k}$, and we see that a vector perpendicular to that is $-\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j} + 0\,\mathbf{k}$. Since this vector is also a unit vector and points in the (positive) θ direction, it must be $\mathbf{e}_\theta$:

$$\boxed{\mathbf{e}_\theta \;=\; -\sin\theta\,\mathbf{i} \;+\; \cos\theta\,\mathbf{j} \;+\; 0\,\mathbf{k}}$$

Lastly, since $\mathbf{e}_\phi = \mathbf{e}_\theta \times \mathbf{e}_\rho$, we get:

$$\boxed{\mathbf{e}_\phi \;=\; \cos\phi\cos\theta\,\mathbf{i} \;+\; \cos\phi\sin\theta\,\mathbf{j} \;-\; \sin\phi\,\mathbf{k}}$$

Step 2: Use the three formulas from Step 1 to solve for $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ in terms of $\mathbf{e}_\rho$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$.

This comes down to solving a system of three equations in three unknowns. There are many ways of doing this, but we will do it by combining the formulas for $\mathbf{e}_\rho$ and $\mathbf{e}_\phi$ to eliminate $\mathbf{k}$, which will give us an equation involving just $\mathbf{i}$ and $\mathbf{j}$. This, with the formula for $\mathbf{e}_\theta$, will then leave us with a system of two equations in two unknowns ($\mathbf{i}$ and $\mathbf{j}$), which we will use to solve first for $\mathbf{j}$ then for $\mathbf{i}$. Lastly, we will solve for $\mathbf{k}$.

First, note that

$$\sin\phi\,\mathbf{e}_\rho \;+\; \cos\phi\,\mathbf{e}_\phi \;=\; \cos\theta\,\mathbf{i} \;+\; \sin\theta\,\mathbf{j}$$

so that

$$\sin\theta(\sin\phi\,\mathbf{e}_\rho + \cos\phi\,\mathbf{e}_\phi) + \cos\theta\,\mathbf{e}_\theta = (\sin^2\theta + \cos^2\theta)\mathbf{j} = \mathbf{j}\,,$$

and so:

$$\boxed{\mathbf{j} = \sin\phi\sin\theta\,\mathbf{e}_\rho + \cos\theta\,\mathbf{e}_\theta + \cos\phi\sin\theta\,\mathbf{e}_\phi}$$

Likewise, we see that

$$\cos\theta(\sin\phi\,\mathbf{e}_\rho + \cos\phi\,\mathbf{e}_\phi) - \sin\theta\,\mathbf{e}_\theta = (\cos^2\theta + \sin^2\theta)\mathbf{i} = \mathbf{i}\,,$$

and so:

$$\boxed{\mathbf{i} = \sin\phi\cos\theta\,\mathbf{e}_\rho - \sin\theta\,\mathbf{e}_\theta + \cos\phi\cos\theta\,\mathbf{e}_\phi}$$

Lastly, we see that:

$$\boxed{\mathbf{k} = \cos\phi\,\mathbf{e}_\rho - \sin\phi\,\mathbf{e}_\phi}$$

Step 3: Get formulas for $\frac{\partial F}{\partial \rho}$, $\frac{\partial F}{\partial \theta}$, $\frac{\partial F}{\partial \phi}$ in terms of $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, $\frac{\partial F}{\partial z}$.

By the Chain Rule, we have

$$\begin{aligned}
\frac{\partial F}{\partial \rho} &= \frac{\partial F}{\partial x}\frac{\partial x}{\partial \rho} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial \rho} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial \rho}\,,\\
\frac{\partial F}{\partial \theta} &= \frac{\partial F}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial \theta} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial \theta}\,,\\
\frac{\partial F}{\partial \phi} &= \frac{\partial F}{\partial x}\frac{\partial x}{\partial \phi} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial \phi} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial \phi}\,,
\end{aligned}$$

which yields:

$$\boxed{\begin{aligned}
\frac{\partial F}{\partial \rho} &= \sin\phi\cos\theta\,\frac{\partial F}{\partial x} + \sin\phi\sin\theta\,\frac{\partial F}{\partial y} + \cos\phi\,\frac{\partial F}{\partial z}\\
\frac{\partial F}{\partial \theta} &= -\rho\sin\phi\sin\theta\,\frac{\partial F}{\partial x} + \rho\sin\phi\cos\theta\,\frac{\partial F}{\partial y}\\
\frac{\partial F}{\partial \phi} &= \rho\cos\phi\cos\theta\,\frac{\partial F}{\partial x} + \rho\cos\phi\sin\theta\,\frac{\partial F}{\partial y} - \rho\sin\phi\,\frac{\partial F}{\partial z}
\end{aligned}}$$

Step 4: Use the three formulas from Step 3 to solve for $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, $\frac{\partial F}{\partial z}$ in terms of $\frac{\partial F}{\partial \rho}$, $\frac{\partial F}{\partial \theta}$, $\frac{\partial F}{\partial \phi}$.

Again, this involves solving a system of three equations in three unknowns. Using a similar process of elimination as in Step 2, we get:

$$\boxed{\begin{aligned}
\frac{\partial F}{\partial x} &= \frac{1}{\rho\sin\phi}\left(\rho\sin^2\phi\cos\theta\,\frac{\partial F}{\partial \rho} - \sin\theta\,\frac{\partial F}{\partial \theta} + \sin\phi\cos\phi\cos\theta\,\frac{\partial F}{\partial \phi}\right)\\
\frac{\partial F}{\partial y} &= \frac{1}{\rho\sin\phi}\left(\rho\sin^2\phi\sin\theta\,\frac{\partial F}{\partial \rho} + \cos\theta\,\frac{\partial F}{\partial \theta} + \sin\phi\cos\phi\sin\theta\,\frac{\partial F}{\partial \phi}\right)\\
\frac{\partial F}{\partial z} &= \frac{1}{\rho}\left(\rho\cos\phi\,\frac{\partial F}{\partial \rho} - \sin\phi\,\frac{\partial F}{\partial \phi}\right)
\end{aligned}}$$

Step 5: Substitute the formulas for **i**, **j**, **k** from Step 2 and the formulas for $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, $\frac{\partial F}{\partial z}$ from Step 4 into the Cartesian gradient formula $\nabla F(x,y,z) = \frac{\partial F}{\partial x}\,\mathbf{i} + \frac{\partial F}{\partial y}\,\mathbf{j} + \frac{\partial F}{\partial z}\,\mathbf{k}$.

Doing this last step is perhaps the most tedious, since it involves simplifying $3\times 3+3\times 3+2\times 2 = 22$ terms! Namely,

$$\begin{aligned}\nabla F \;=\; & \frac{1}{\rho\sin\phi}\left(\rho\sin^2\phi\cos\theta\,\frac{\partial F}{\partial \rho} - \sin\theta\,\frac{\partial F}{\partial \theta} + \sin\phi\cos\phi\cos\theta\,\frac{\partial F}{\partial \phi}\right)(\sin\phi\cos\theta\,\mathbf{e}_\rho - \sin\theta\,\mathbf{e}_\theta \\ & + \cos\phi\cos\theta\,\mathbf{e}_\phi) \\ + \; & \frac{1}{\rho\sin\phi}\left(\rho\sin^2\phi\sin\theta\,\frac{\partial F}{\partial \rho} + \cos\theta\,\frac{\partial F}{\partial \theta} + \sin\phi\cos\phi\sin\theta\,\frac{\partial F}{\partial \phi}\right)(\sin\phi\sin\theta\,\mathbf{e}_\rho + \cos\theta\,\mathbf{e}_\theta \\ & + \cos\phi\sin\theta\,\mathbf{e}_\phi) \\ + \; & \frac{1}{\rho}\left(\rho\cos\phi\,\frac{\partial F}{\partial \rho} - \sin\phi\,\frac{\partial F}{\partial \phi}\right)(\cos\phi\,\mathbf{e}_\rho - \sin\phi\,\mathbf{e}_\phi)\,,\end{aligned}$$

which we see has 8 terms involving $\mathbf{e}_\rho$, 6 terms involving $\mathbf{e}_\theta$, and 8 terms involving $\mathbf{e}_\phi$. But the algebra is straightforward and yields the desired result:

$$\nabla F \;=\; \frac{\partial F}{\partial \rho}\,\mathbf{e}_\rho \;+\; \frac{1}{\rho\sin\phi}\,\frac{\partial F}{\partial \theta}\,\mathbf{e}_\theta \;+\; \frac{1}{\rho}\,\frac{\partial F}{\partial \phi}\,\mathbf{e}_\phi \qquad \checkmark$$

Example 4.19. In Example 4.17 we showed that $\nabla\|\mathbf{r}\|^2 = 2\,\mathbf{r}$ and $\Delta\|\mathbf{r}\|^2 = 6$, where $\mathbf{r}(x,y,z) = x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}$ in Cartesian coordinates. Verify that we get the same answers if we switch to spherical coordinates.

Solution: Since $\|\mathbf{r}\|^2 = x^2+y^2+z^2 = \rho^2$ in spherical coordinates, let $F(\rho,\theta,\phi) = \rho^2$ (so that $F(\rho,\theta,\phi) = \|\mathbf{r}\|^2$). The gradient of F in spherical coordinates is

$$\begin{aligned}\nabla F \;&=\; \frac{\partial F}{\partial \rho}\,\mathbf{e}_\rho \;+\; \frac{1}{\rho\sin\phi}\,\frac{\partial F}{\partial \theta}\,\mathbf{e}_\theta \;+\; \frac{1}{\rho}\,\frac{\partial F}{\partial \phi}\,\mathbf{e}_\phi \\ &=\; 2\rho\,\mathbf{e}_\rho \;+\; \frac{1}{\rho\sin\phi}(0)\,\mathbf{e}_\theta \;+\; \frac{1}{\rho}(0)\,\mathbf{e}_\phi \\ &=\; 2\rho\,\mathbf{e}_\rho \;=\; 2\rho\,\frac{\mathbf{r}}{\|\mathbf{r}\|}\;, \text{ as we showed earlier, so} \\ &=\; 2\rho\,\frac{\mathbf{r}}{\rho} \;=\; 2\,\mathbf{r}\;, \text{ as expected. And the Laplacian is} \\ \Delta F \;&=\; \frac{1}{\rho^2}\,\frac{\partial}{\partial \rho}\left(\rho^2\,\frac{\partial F}{\partial \rho}\right) \;+\; \frac{1}{\rho^2\sin^2\phi}\,\frac{\partial^2 F}{\partial \theta^2} \;+\; \frac{1}{\rho^2\sin\phi}\,\frac{\partial}{\partial \phi}\left(\sin\phi\,\frac{\partial F}{\partial \phi}\right) \\ &=\; \frac{1}{\rho^2}\,\frac{\partial}{\partial \rho}(\rho^2\,2\rho) \;+\; \frac{1}{\rho^2\sin\phi}(0) \;+\; \frac{1}{\rho^2\sin\phi}\,\frac{\partial}{\partial \phi}\big(\sin\phi(0)\big) \\ &=\; \frac{1}{\rho^2}\,\frac{\partial}{\partial \rho}(2\rho^3) \;+\; 0 \;+\; 0 \\ &=\; \frac{1}{\rho^2}(6\rho^2) \;=\; 6\;, \text{ as expected.}\end{aligned}$$

Exercises

A

For Exercises 1-6, find the Laplacian of the function $f(x,y,z)$ in Cartesian coordinates.

1. $f(x,y,z) = x + y + z$ **2.** $f(x,y,z) = x^5$ **3.** $f(x,y,z) = (x^2 + y^2 + z^2)^{3/2}$

4. $f(x,y,z) = e^{x+y+z}$ **5.** $f(x,y,z) = x^3 + y^3 + z^3$ **6.** $f(x,y,z) = e^{-x^2-y^2-z^2}$

7. Find the Laplacian of the function in Exercise 3 in spherical coordinates.

8. Find the Laplacian of the function in Exercise 6 in spherical coordinates.

9. Let $f(x,y,z) = \dfrac{z}{x^2+y^2}$ in Cartesian coordinates. Find ∇f in cylindrical coordinates.

10. For $\mathbf{f}(r,\theta,z) = r\,\mathbf{e}_r + z\sin\theta\,\mathbf{e}_\theta + rz\,\mathbf{e}_z$ in cylindrical coordinates, find div $\mathbf{f}$ and curl $\mathbf{f}$.

11. For $\mathbf{f}(\rho,\theta,\phi) = \mathbf{e}_\rho + \rho\cos\theta\,\mathbf{e}_\theta + \rho\,\mathbf{e}_\phi$ in spherical coordinates, find div $\mathbf{f}$ and curl $\mathbf{f}$.

B

For Exercises 12-23, prove the given formula ($r = \|\mathbf{r}\|$ is the length of the position vector field $\mathbf{r}(x,y,z) = x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}$).

12. $\nabla(1/r) = -\mathbf{r}/r^3$ **13.** $\Delta(1/r) = 0$ **14.** $\nabla\cdot(\mathbf{r}/r^3) = 0$ **15.** $\nabla(\ln r) = \mathbf{r}/r^2$

16. $\text{div}(\mathbf{F}+\mathbf{G}) = \text{div}\,\mathbf{F} + \text{div}\,\mathbf{G}$

17. $\text{curl}(\mathbf{F}+\mathbf{G}) = \text{curl}\,\mathbf{F} + \text{curl}\,\mathbf{G}$

18. $\text{div}(f\,\mathbf{F}) = f\,\text{div}\,\mathbf{F} + \mathbf{F}\cdot\nabla f$

19. $\text{div}(\mathbf{F}\times\mathbf{G}) = \mathbf{G}\cdot\text{curl}\,\mathbf{F} - \mathbf{F}\cdot\text{curl}\,\mathbf{G}$

20. $\text{div}(\nabla f\times\nabla g) = 0$

21. $\text{curl}(f\,\mathbf{F}) = f\,\text{curl}\,\mathbf{F} + (\nabla f)\times\mathbf{F}$

22. $\text{curl}(\text{curl}\,\mathbf{F}) = \nabla(\text{div}\,\mathbf{F}) - \Delta\mathbf{F}$

23. $\Delta(fg) = f\,\Delta g + g\,\Delta f + 2(\nabla f\cdot\nabla g)$

C

24. Prove Theorem 4.17.

25. Derive the gradient formula in cylindrical coordinates: $\nabla F = \frac{\partial F}{\partial r}\,\mathbf{e}_r + \frac{1}{r}\,\frac{\partial F}{\partial \theta}\,\mathbf{e}_\theta + \frac{\partial F}{\partial z}\,\mathbf{e}_z$

26. Use $\mathbf{f} = u\,\nabla v$ in the Divergence Theorem to prove:

(a) *Green's first identity:* $\displaystyle\iiint\limits_S (u\,\Delta v + (\nabla u)\cdot(\nabla v))\,dV = \iint\limits_\Sigma (u\,\nabla v)\cdot d\boldsymbol{\sigma}$

(b) *Green's second identity:* $\displaystyle\iiint\limits_S (u\,\Delta v - v\,\Delta u)\,dV = \iint\limits_\Sigma (u\,\nabla v - v\,\nabla u)\cdot d\boldsymbol{\sigma}$

27. Suppose that $\Delta u = 0$ (i.e. u is *harmonic*) over $\mathbb{R}^3$. Define the *normal derivative* $\frac{\partial u}{\partial n}$ of u over a closed surface Σ with outward unit normal vector $\mathbf{n}$ by $\frac{\partial u}{\partial n} = D_{\mathbf{n}}u = \mathbf{n}\cdot\nabla u$. Show that $\displaystyle\iint\limits_\Sigma \frac{\partial u}{\partial n}\,d\sigma = 0$. (*Hint: Use Green's second identity.*)

Bibliography

Abbott, E.A., *Flatland*, 7th edition. New York: Dover Publications, Inc., 1952

Classic tale about a creature living in a 2-dimensional world who encounters a higher-dimensional creature, with lots of humor thrown in.

Anton, H. and C. Rorres, *Elementary Linear Algebra: Applications Version*, 8th edition. New York: John Wiley & Sons, 2000

Standard treatment of elementary linear algebra.

Bazaraa, M.S., H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd edition. New York: John Wiley & Sons, 1993

Thorough treatment of nonlinear optimization.

Farin, G., *Curves and Surfaces for Computer Aided Geometric Design: A Practical Guide*, 2nd edition. San Diego, CA: Academic Press, 1990

An intermediate-level book on curve and surface design.

Hecht, E., *Optics*, 2nd edition. Reading, MA: Addison-Wesley Publishing Co., 1987

An intermediate-level book on optics, covering a wide range of topics.

Hoel, P.G., S.C. Port and C.J. Stone, *Introduction to Probability Theory*, Boston, MA: Houghton Mifflin Co., 1971

An excellent introduction to elementary, calculus-based probability theory. Lots of good exercises.

Jackson, J.D., *Classical Electrodynamics*, 2nd edition. New York: John Wiley & Sons, 1975

An advanced book on electromagnetism, famous for being intimidating. Most of the mathematics will be understandable after reading the present book.

Marion, J.B., *Classical Dynamics of Particles and Systems*, 2nd edition. New York: Academic Press, 1970

Standard intermediate-level treatment of classical mechanics. Very thorough.

O'Neill, B., *Elementary Differential Geometry*, New York: Academic Press, 1966

Intermediate-level book on differential geometry, with a modern approach based on differential forms.

Pogorelov, A.V., *Analytical Geometry*, Moscow: Mir Publishers, 1980

An intermediate/advanced book on analytic geometry.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd edition. Cambridge, UK: Cambridge University Press, 1992

An excellent source of information on numerical methods for solving a wide variety of problems. Though all the examples are in the `FORTRAN` programming language, the code is clear enough to implement in the language of your choice.

Protter, M.H. and C.B. Morrey, *Analytic Geometry*, 2nd edition. Reading, MA: Addison-Wesley Publishing Co., 1975

Thorough treatment of elementary analytic geometry, with a rigor not found in most recent books.

Ralston, A. and P. Rabinowitz, *A First Course in Numerical Analysis*, 2nd edition. New York: McGraw-Hill, 1978

Standard treatment of elementary numerical analysis.

Reitz, J.R., F.J. Milford and R.W. Christy, *Foundations of Electromagnetic Theory*, 3rd edition. Reading, MA: Addison-Wesley Publishing Co., 1979

Intermediate text on electromagnetism.

Schey, H.M., *Div, Grad, Curl, and All That: An Informal Text on Vector Calculus*, New York: W.W. Norton & Co., 1973

Very intuitive approach to the subject, from a physicist's viewpoint. Highly recommended.

Taylor, A.E. and W.R. Mann, *Advanced Calculus*, 2nd edition. New York: John Wiley & Sons, 1972

Excellent treatment of n-dimensional calculus. A good book to study after the present book. Many intriguing exercises.

Uspensky, J.V., *Theory of Equations*, New York: McGraw-Hill, 1948

A classic on the subject, discussing many interesting topics.

Weinberger, H.F., *A First Course in Partial Differential Equations*, New York: John Wiley & Sons, 1965

A good introduction to the vast subject of partial differential equations.

Welchons, A.M. and W.R. Krickenberger, *Solid Geometry*, Boston, MA: Ginn & Co., 1936

A very thorough treatment of 3-dimensional geometry from an elementary perspective, includes many topics which (sadly) do not seem to be taught anymore.

Appendix A

Answers and Hints to Selected Exercises

Chapter 1

Section 1.1 (p. 8)

1. (a) $\sqrt{5}$ (b) $\sqrt{5}$ (c) $\sqrt{17}$ (d) 1 (e) $2\sqrt{17}$ **2.** Yes **3.** No

Section 1.2 (p. 14)

1. (a) $(-4,4,-3)$ (b) $(2,6,-1)$ (c) $\left(\frac{-1}{\sqrt{30}},\frac{5}{\sqrt{30}},\frac{-2}{\sqrt{30}}\right)$ (d) $\frac{\sqrt{41}}{2}$ (e) $\frac{\sqrt{41}}{2}$ (f) $(14,-6,8)$ (g) $(-7,3,-4)$ (h) $(-1,-6,1)$ (i) $(-2,-4,2)$ (j) No.
3. No. $\|\mathbf{v}\|+\|\mathbf{w}\|$ is larger.

Section 1.3 (p. 18)

1. 10 **3.** 73.4° **5.** 90° **7.** 0°
9. Yes, since $\mathbf{v}\cdot\mathbf{w}=0$.
11. $|\mathbf{v}\cdot\mathbf{w}|=0<\sqrt{21}\sqrt{5}=\|\mathbf{v}\|\,\|\mathbf{w}\|$
13. $\|\mathbf{v}+\mathbf{w}\|=\sqrt{26}<\sqrt{21}+\sqrt{5}=\|\mathbf{v}\|+\|\mathbf{w}\|$
15. Hint: use Definition 1.6.
24. Hint: See Theorem 1.10(c).

Section 1.4 (p. 29)

1. $(-5,-23,-24)$ **3.** $(8,4,-5)$ **5.** **0** **7.** 16.72 **9.** $4\sqrt{5}$ **11.** 9 **13.** 0 and $(8,-10,2)$ **15.** 14

Section 1.5 (p. 39)

1. (a) $(2,3,-2)+t(5,4,-3)$ (b) $x=2+5t$, $y=3+4t$, $z=-2-3t$ (c) $\frac{x-2}{5}=\frac{y-3}{4}=\frac{z+2}{-3}$
3. (a) $(2,1,3)+t(1,0,1)$ (b) $x=2+t$, $y=1$, $z=3+t$ (c) $x-2=z-3$, $y=1$
5. $x=1+2t$, $y=-2+7t$, $z=-3+8t$
7. 7.65 **9.** $(1,2,3)$
11. $4x-4y+3z-10=0$
13. $x-2y-z+2=0$
15. $11x-24y+21z-26=0$ **17.** $9/\sqrt{35}$
19. $x=5t$, $y=2+3t$, $z=-7t$ **21.** $(10,-2,1)$

Section 1.6 (p. 46)

1. radius: 1, center: $(2,3,5)$ **3.** radius: 5, center: $(-1,-1,-1)$ **5.** No intersection.
7. circle $x^2+y^2=4$ in the planes $z=\pm\sqrt{5}$
9. lines $\frac{x}{a}=\frac{y}{b}$, $z=0$ and $\frac{x}{a}=-\frac{y}{b}$, $z=0$
13. $\left(\frac{2a}{2-c},\frac{2b}{2-c},0\right)$

Section 1.7 (p. 50)

1. (a) $(4,\frac{\pi}{3},-1)$ (b) $(\sqrt{17},\frac{\pi}{3},1.816)$
3. (a) $(2\sqrt{7},\frac{11\pi}{6},0)$ (b) $(2\sqrt{7},\frac{11\pi}{6},\frac{\pi}{2})$
5. (a) $r^2+z^2=25$ (b) $\rho=5$
7. (a) $r^2+9z^2=36$ (b) $\rho^2(1+8\cos^2\phi)=36$
10. $(a,\theta,a\cot\phi)$ **12.** Hint: Use the distance formula for Cartesian coordinates.

Section 1.8 (p. 57)

1. $\mathbf{f}'(t)=(1,2t,3t^2)$; $x=1+t$, $y=z=1$
3. $\mathbf{f}'(t)=(-2\sin 2t,2\cos 2t,1)$; $x=1$, $y=2t$, $z=t$ **5.** $\mathbf{v}(t)=(1,1-\cos t,\sin t)$, $\mathbf{a}(t)=(0,\sin t,\cos t)$
9. (a) Line parallel to $\mathbf{c}$ (b) Half-line parallel to $\mathbf{c}$ (c) Hint: Think of the

functions as position vectors.
15. Hint: Theorem 1.16

Section 1.9 (p. 63)

1. $\frac{3\pi\sqrt{5}}{2}$ **3.** $2(5^{3/2}-8)$ **5.** Replace t by $\left(\left(\frac{27s+16}{2}\right)^{2/3}-4\right)\Big/9$ **6.** Hint: Use Theorem 1.20(e), Example 1.37, and Theorem 1.16 **7.** Hint: Use Exercise 6.
9. Hint: Use $\mathbf{f}'(t) = \|\mathbf{f}(t)\|\mathbf{T}$, differentiate that to get $\mathbf{f}''(t)$, put those expressions into $\mathbf{f}'(t) \times \mathbf{f}''(t)$, then write $\mathbf{T}'(t)$ in terms of $\mathbf{N}(t)$. **11.** $\mathbf{T}(t) = \frac{1}{\sqrt{2}}(-\sin t, \cos t, 1)$, $\mathbf{N}(t) = (-\cos t, -\sin t, 0)$, $\mathbf{B}(t) = \frac{1}{\sqrt{2}}(\sin t, -\cos t, 1)$, $\kappa(t) = 1/2$

Chapter 2

Section 2.1 (p. 70)

1. domain: $\mathbb{R}^2$, range: $[-1,\infty)$ **3.** domain: $\{(x,y): x^2+y^2 \ge 4\}$, range: $[0,\infty)$
5. domain: $\mathbb{R}^3$, range: $[-1,1]$ **7.** 1
9. does not exist **11.** 2 **13.** 2 **15.** 0
17. does not exist

Section 2.2 (p. 74)

1. $\frac{\partial f}{\partial x} = 2x$, $\frac{\partial f}{\partial y} = 2y$ **3.** $\frac{\partial f}{\partial x} = x(x^2+y+4)^{-1/2}$, $\frac{\partial f}{\partial y} = \frac{1}{2}(x^2+y+4)^{-1/2}$ **5.** $\frac{\partial f}{\partial x} = ye^{xy}+y$, $\frac{\partial f}{\partial y} = xe^{xy}+x$ **7.** $\frac{\partial f}{\partial x} = 4x^3$, $\frac{\partial f}{\partial y} = 0$
9. $\frac{\partial f}{\partial x} = x(x^2+y^2)^{-1/2}$, $\frac{\partial f}{\partial y} = y(x^2+y^2)^{-1/2}$
11. $\frac{\partial f}{\partial x} = \frac{2x}{3}(x^2+y+4)^{-2/3}$, $\frac{\partial f}{\partial y} = \frac{1}{3}(x^2+y+4)^{-2/3}$ **13.** $\frac{\partial f}{\partial x} = -2xe^{-(x^2+y^2)}$, $\frac{\partial f}{\partial y} = -2ye^{-(x^2+y^2)}$ **15.** $\frac{\partial f}{\partial x} = y\cos(xy)$, $\frac{\partial f}{\partial y} = x\cos(xy)$ **17.** $\frac{\partial^2 f}{\partial x^2} = 2$, $\frac{\partial^2 f}{\partial y^2} = 2$, $\frac{\partial^2 f}{\partial x\,\partial y} = 0$ **19.** $\frac{\partial^2 f}{\partial x^2} = (y+4)(x^2+y+4)^{-3/2}$, $\frac{\partial^2 f}{\partial y^2} = -\frac{1}{4}(x^2+y+4)^{-3/2}$, $\frac{\partial^2 f}{\partial x\,\partial y} = -\frac{1}{2}x(x^2+y+4)^{-3/2}$
21. $\frac{\partial^2 f}{\partial x^2} = y^2e^{xy}$, $\frac{\partial^2 f}{\partial y^2} = x^2e^{xy}$, $\frac{\partial^2 f}{\partial x\,\partial y} = (1+xy)e^{xy}+1$ **23.** $\frac{\partial^2 f}{\partial x^2} = 12x^2$, $\frac{\partial^2 f}{\partial y^2} = 0$, $\frac{\partial^2 f}{\partial x\,\partial y} = 0$ **25.** $\frac{\partial^2 f}{\partial x^2} = -x^{-2}$, $\frac{\partial^2 f}{\partial y^2} = -y^{-2}$, $\frac{\partial^2 f}{\partial x\,\partial y} = 0$

Section 2.3 (p. 77)

1. $2x+3y-z-3=0$ **3.** $-2x+y-z-2=0$
5. $x+2y=z$ **7.** $\frac{1}{2}(x-1)+\frac{4}{9}(y-2)+\frac{\sqrt{11}}{12}(z-\frac{2\sqrt{11}}{3})=0$ **9.** $3x+4y-5z=0$

Section 2.4 (p. 82)

1. $(2x,2y)$ **3.** $(\frac{x}{\sqrt{x^2+y^2+4}}, \frac{y}{\sqrt{x^2+y^2+4}})$
5. $(1/x, 1/y)$ **7.** $(yz\cos(xyz), xz\cos(xyz), xy\cos(xyz))$
9. $(2x,2y,2z)$ **11.** $2\sqrt{2}$ **13.** $\frac{1}{\sqrt{3}}$
15. $\sqrt{3}\cos(1)$ **17.** increase: $(45,20)$, decrease: $(-45,-20)$

Section 2.5 (p. 88)

1. local min. $(1,0)$; saddle pt. $(-1,0)$
3. local min. $(1,1)$; local max. $(-1,-1)$; saddle pts. $(1,-1),(-1,1)$ **5.** local min. $(1,-1)$; saddle pt. $(0,0)$ **7.** local min. $(0,0)$
9. local min. $(-1,1/2)$ **11.** width = height = depth=10 **13.** $x=y=4$, $z=2$

Section 2.6 (p. 95)

2. $(x_0,y_0) = (0,0): \to (0.2858, -0.3998)$; $(x_0,y_0) = (1,1): \to (1.03256, -1.94037)$

Section 2.7 (p. 100)

1. min. $\left(\frac{-4}{\sqrt{5}}, \frac{-2}{\sqrt{5}}\right)$; max. $\left(\frac{4}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right)$
3. min. $\left(\frac{20}{\sqrt{13}}, \frac{30}{\sqrt{13}}\right)$; max. $\left(-\frac{20}{\sqrt{13}}, -\frac{30}{\sqrt{13}}\right)$
4. min. $\left(\frac{-9}{\sqrt{5}}, 0, \frac{2}{\sqrt{5}}\right)$; max. $\left(\frac{9}{8}, \frac{\sqrt{59}}{4}, \frac{-1}{4}\right)$
5. $\frac{8abc}{3\sqrt{3}}$

Chapter 3

Section 3.1 (p. 104)

1. 1 **3.** $\frac{7}{12}$ **5.** $\frac{7}{6}$ **7.** 5 **9.** $\frac{1}{2}$ **11.** 15

Section 3.2 (p. 109)

1. 1 **3.** $8\ln 2 - 3$ **5.** $\frac{\pi}{4}$ **6.** $\frac{1}{4}$ **7.** 2 **9.** $\frac{1}{6}$
10. $\frac{6}{5}$

Section 3.3 (p. 112)

1. $\frac{9}{2}$ **3.** $(2\cos(\pi^2) + \pi^4 - 2)/4$ **5.** $\frac{1}{6}$ **7.** 6
10. $\frac{1}{3}$

Section 3.4 (p. 116)

1. The values should converge to ≈ 1.318. (Hint: In Java the exponential function e^x can be obtained with `Math.exp(x)`. Other languages have similar functions, otherwise use $e = 2.7182818284590455$ in your program.)
2. ≈ 1.146 **3.** ≈ 0.705 **4.** ≈ 0.168

Section 3.5 (p. 123)

1. 8π **3.** $\frac{4\pi}{3}(8 - 3^{3/2})$ **7.** $1 - \frac{\sin 2}{2}$ **9.** $2\pi ab$

Section 3.6 (p. 127)

1. $(1, 8/3)$ **3.** $(0, \frac{4a}{3\pi})$ **5.** $(0, 3\pi/16)$
7. $(0, 0, 5a/12)$ **9.** $(7/12, 7/12, 7/12)$

Section 3.7 (p. 134)

1. $\sqrt{\pi}$ **2.** 1 **6.** Both are $\frac{n}{(n+1)^2(n+2)}$ **7.** $\frac{1}{n}$

Chapter 4

Section 4.1 (p. 142)

1. 1/2 **3.** 23 **5.** 24π **7.** -2π **9.** 2π
11. 0

Section 4.2 (p. 149)

1. 0 **3.** No **4.** Yes. $F(x,y) = \frac{x^2}{2} - \frac{y^2}{2}$
5. No **9.** (b) No. Hint: Think of how F is defined. **10.** Yes. $F(x,y) = axy + bx + cy + d$

Section 4.3 (p. 155)

1. 16/15 **3.** -5π **5.** Yes. $F(x,y) = xy^2 + x^3$
7. Yes. $F(x,y) = 4x^2y + 2y^2 + 3x$

Section 4.4 (p. 163)

1. 216π **2.** 3 **3.** $12\pi/5$ **7.** 15/4

Section 4.5 (p. 175)

1. $2\sqrt{2}\pi^2$ **2.** $(17\sqrt{17} - 5\sqrt{5})/3$ **3.** 2/5
4. 2 **5.** $2\pi(\pi - 1)$ **7.** 67/15 **9.** 6
11. Yes **13.** No **19.** Hint: Think of how a vector field $\mathbf{f}(x,y) = P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$ in $\mathbb{R}^2$ can be extended in a natural way to be a vector field in $\mathbb{R}^3$.

Section 4.6 (p. 186)

1. 0 **3.** $12\sqrt{x^2 + y^2 + z^2}$ **5.** $6(x + y + z)$
7. 12ρ **8.** $(4\rho^2 - 6)e^{-\rho^2}$ **9.** $-\frac{2z}{r^3}\,\mathbf{e}_r + \frac{1}{r^2}\,\mathbf{e}_z$
11. div $\mathbf{f} = \frac{2}{\rho} - \frac{\sin\theta}{\sin\phi} + \cot\phi$;
curl $\mathbf{f} = \cot\phi\cos\theta\,\mathbf{e}_\rho + 2\mathbf{e}_\theta - 2\cos\theta\,\mathbf{e}_\phi$
25. Hint: Start by showing that $\mathbf{e}_r = \cos\theta\,\mathbf{i} + \sin\theta\,\mathbf{j}$, $\mathbf{e}_\theta = -\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j}$, $\mathbf{e}_z = \mathbf{k}$.

Appendix B

We will prove the right-hand rule for the cross product of two vectors in $\mathbb{R}^3$.

For any vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$, define a new vector, $\mathbf{n}(\mathbf{v},\mathbf{w})$, as follows:

1. If $\mathbf{v}$ and $\mathbf{w}$ are nonzero and not parallel, and θ is the angle between them, then $\mathbf{n}(\mathbf{v},\mathbf{w})$ is the vector in $\mathbb{R}^3$ such that:
   (a) the magnitude of $\mathbf{n}(\mathbf{v},\mathbf{w})$ is $\|\mathbf{v}\|\,\|\mathbf{w}\|\,\sin\theta$,
   (b) $\mathbf{n}(\mathbf{v},\mathbf{w})$ is perpendicular to the plane containing $\mathbf{v}$ and $\mathbf{w}$, and
   (c) $\mathbf{v}$, $\mathbf{w}$, $\mathbf{n}(\mathbf{v},\mathbf{w})$ form a right-handed system.
2. If $\mathbf{v}$ and $\mathbf{w}$ are nonzero and parallel, then $\mathbf{n}(\mathbf{v},\mathbf{w}) = \mathbf{0}$.
3. If either $\mathbf{v}$ or $\mathbf{w}$ is $\mathbf{0}$, then $\mathbf{n}(\mathbf{v},\mathbf{w}) = \mathbf{0}$.

The goal is to show that $\mathbf{n}(\mathbf{v},\mathbf{w}) = \mathbf{v}\times\mathbf{w}$ for all $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^3$, which would prove the right-hand rule for the cross product (by part 1(c) of our definition). To do this, we will perform the following steps:

Step 1: Show that $\mathbf{n}(\mathbf{v},\mathbf{w}) = \mathbf{v}\times\mathbf{w}$ if $\mathbf{v}$ and $\mathbf{w}$ are any two of the basis vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$.
This was already shown in Example 1.11 in Section 1.4. ✓

Step 2: Show that $\mathbf{n}(a\mathbf{v}, b\mathbf{w}) = ab(\mathbf{v}\times\mathbf{w})$ for any scalars a, b if $\mathbf{v}$ and $\mathbf{w}$ are any two of the basis vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$.
If either $a = 0$ or $b = 0$ then $\mathbf{n}(a\mathbf{v}, b\mathbf{w}) = \mathbf{0} = ab(\mathbf{v}\times\mathbf{w})$, so the result holds. So assume that $a \neq 0$ and $b \neq 0$. Let $\mathbf{v}$ and $\mathbf{w}$ be any two of the basis vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$. For example, we will show that the result holds for $\mathbf{v} = \mathbf{i}$ and $\mathbf{w} = \mathbf{k}$ (the other possibilities follow in a similar fashion).

For $a\mathbf{v} = a\mathbf{i}$ and $b\mathbf{w} = b\mathbf{k}$, the angle θ between $a\mathbf{v}$ and $b\mathbf{w}$ is 90°. Hence the magnitude of $\mathbf{n}(a\mathbf{v}, b\mathbf{w})$, by definition, is $\|a\mathbf{i}\|\,\|b\mathbf{k}\|\,\sin 90° = |ab|$. Also, by definition, $\mathbf{n}(a\mathbf{v}, b\mathbf{w})$ is perpendicular to the plane containing $a\mathbf{i}$ and $b\mathbf{k}$, namely, the xz-plane. Thus, $\mathbf{n}(a\mathbf{v}, b\mathbf{w})$ must be a scalar multiple of $\mathbf{j}$. Since its magnitude is $|ab|$, then $\mathbf{n}(a\mathbf{v}, b\mathbf{w})$ must be either $|ab|\mathbf{j}$ or $-|ab|\mathbf{j}$.

There are four possibilities for the combinations of signs for a and b. We will consider the case when $a > 0$ and $b > 0$ (the other three possibilities are handled similarly).

In this case, $\mathbf{n}(a\mathbf{v}, b\mathbf{w})$ must be either $ab\mathbf{j}$ or $-ab\mathbf{j}$. Now, since $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ form a right-handed system, then $\mathbf{i}$, $\mathbf{k}$, $\mathbf{j}$ form a left-handed system, and so $\mathbf{i}$, $\mathbf{k}$, $-\mathbf{j}$ form a right-handed system. Thus, $a\mathbf{i}$, $b\mathbf{k}$, $-ab\mathbf{j}$ form a right-handed system (since $a > 0$, $b > 0$, and $ab > 0$). So since, by definition, $a\mathbf{i}$, $b\mathbf{k}$, $\mathbf{n}(a\mathbf{i}, b\mathbf{k})$ form a right-handed system, and since $\mathbf{n}(a\mathbf{i}, b\mathbf{k})$ has to be either $ab\mathbf{j}$ or $-ab\mathbf{j}$, this means that we must have $\mathbf{n}(a\mathbf{i}, b\mathbf{k}) = -ab\mathbf{j}$.

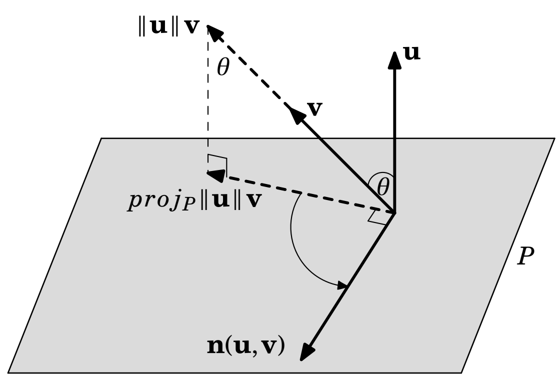
But we know that $a\mathbf{i} \times b\mathbf{k} = ab(\mathbf{i} \times \mathbf{k}) = ab(-\mathbf{j}) = -ab\mathbf{j}$. Therefore, $\mathbf{n}(a\mathbf{i}, b\mathbf{k}) = ab(\mathbf{i} \times \mathbf{k})$, which is what we needed to show.

$\therefore\ \mathbf{n}(a\mathbf{v}, b\mathbf{w}) = ab(\mathbf{v} \times \mathbf{w})$ ✓

Step 3: Show that $\mathbf{n}(\mathbf{u}, \mathbf{v}+\mathbf{w}) = \mathbf{n}(\mathbf{u}, \mathbf{v}) + \mathbf{n}(\mathbf{u}, \mathbf{w})$ for any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$.

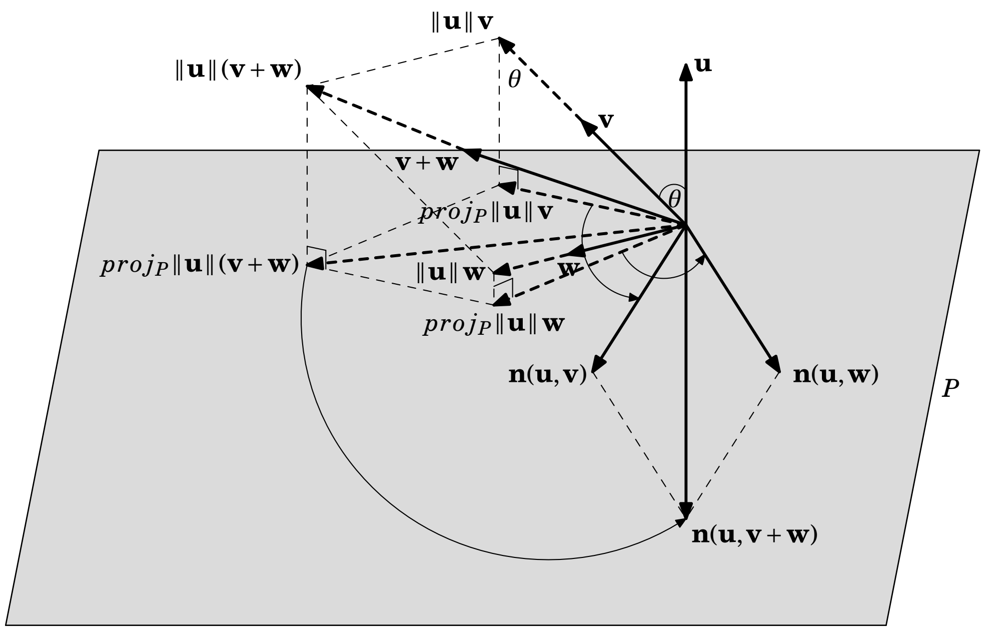
If $\mathbf{u} = \mathbf{0}$ then the result holds trivially since $\mathbf{n}(\mathbf{u}, \mathbf{v}+\mathbf{w})$, $\mathbf{n}(\mathbf{u}, \mathbf{v})$ and $\mathbf{n}(\mathbf{u}, \mathbf{w})$ are all the zero vector. If $\mathbf{v} = \mathbf{0}$, then the result follows easily since $\mathbf{n}(\mathbf{u}, \mathbf{v}+\mathbf{w}) = \mathbf{n}(\mathbf{u}, \mathbf{0}+\mathbf{w}) = \mathbf{n}(\mathbf{u}, \mathbf{w}) = \mathbf{0} + \mathbf{n}(\mathbf{u}, \mathbf{w}) = \mathbf{n}(\mathbf{u}, \mathbf{0}) = \mathbf{n}(\mathbf{u}, \mathbf{w}) = \mathbf{n}(\mathbf{u}, \mathbf{v}) + \mathbf{n}(\mathbf{u}, \mathbf{w})$. A similar argument shows that the result holds if $\mathbf{w} = \mathbf{0}$.

So now assume that $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are all nonzero vectors. We will describe a geometric construction of $\mathbf{n}(\mathbf{u}, \mathbf{v})$, which is shown in the figure below. Let P be a plane perpendicular to $\mathbf{u}$. Multiply the vector $\mathbf{v}$ by the positive scalar $\|\mathbf{u}\|$, then project the vector $\|\mathbf{u}\|\mathbf{v}$ straight down onto the plane P. You can think of this projection vector (denoted by $proj_P\|\mathbf{u}\|\mathbf{v}$) as the shadow of the vector $\|\mathbf{u}\|\mathbf{v}$ on the plane P, with the light source directly overhead the terminal point of $\|\mathbf{u}\|\mathbf{v}$. If θ is the angle between $\mathbf{u}$ and $\mathbf{v}$, then we see that $proj_P\|\mathbf{u}\|\mathbf{v}$ has magnitude $\|\mathbf{u}\|\,\|\mathbf{v}\|\sin\theta$, which is the magnitude of $\mathbf{n}(\mathbf{u}, \mathbf{v})$. So rotating $proj_P\|\mathbf{u}\|\mathbf{v}$ by 90° in a counter-clockwise direction in the plane P gives a vector whose magnitude is the same as that of $\mathbf{n}(\mathbf{u}, \mathbf{v})$ and which is perpendicular to $proj_P\|\mathbf{u}\|\mathbf{v}$ (and hence perpendicular to $\mathbf{v}$). Since this vector is in P then it is also perpendicular to $\mathbf{u}$. And we can see that $\mathbf{u}$, $\mathbf{v}$ and this vector form a right-handed system. Hence this vector must be $\mathbf{n}(\mathbf{u}, \mathbf{v})$. Note that this holds even if $\mathbf{u} \parallel \mathbf{v}$, since in that case $\theta = 0°$ and so $\sin\theta = 0$ which means that $\mathbf{n}(\mathbf{u}, \mathbf{v})$ has magnitude 0, which is what we would expect.



Now apply this same geometric construction to get $\mathbf{n}(\mathbf{u}, \mathbf{w})$ and $\mathbf{n}(\mathbf{u}, \mathbf{v}+\mathbf{w})$. Since $\|\mathbf{u}\|(\mathbf{v}+\mathbf{w})$ is the sum of the vectors $\|\mathbf{u}\|\mathbf{v}$ and $\|\mathbf{u}\|\mathbf{w}$, then the projection vector $proj_P\|\mathbf{u}\|(\mathbf{v}+\mathbf{w})$ is the sum of the projection vectors $proj_P\|\mathbf{u}\|\mathbf{v}$ and $proj_P\|\mathbf{u}\|\mathbf{w}$ (to see this, using the shadow

analogy again and the parallelogram rule for vector addition, think of how projecting a parallelogram onto a plane gives you a parallelogram in that plane). So then rotating all three projection vectors by 90° in a counter-clockwise direction in the plane P preserves that sum (see the figure below), which means that $\mathbf{n}(\mathbf{u},\mathbf{v}+\mathbf{w}) = \mathbf{n}(\mathbf{u},\mathbf{v}) + \mathbf{n}(\mathbf{u},\mathbf{w})$. ✓



Step 4: Show that $\mathbf{n}(\mathbf{w},\mathbf{v}) = -\mathbf{n}(\mathbf{v},\mathbf{w})$ for any vectors $\mathbf{v}$, $\mathbf{w}$.

If $\mathbf{v}$ and $\mathbf{w}$ are nonzero and parallel, or if either is $\mathbf{0}$, then $\mathbf{n}(\mathbf{w},\mathbf{v}) = \mathbf{0} = -\mathbf{n}(\mathbf{v},\mathbf{w})$, so the result holds. So assume that $\mathbf{v}$ and $\mathbf{w}$ are nonzero and not parallel. Then $\mathbf{n}(\mathbf{w},\mathbf{v})$ has magnitude $\|\mathbf{w}\|\,\|\mathbf{v}\|\sin\theta$, which is the same as the magnitude of $\mathbf{n}(\mathbf{v},\mathbf{w})$, and hence is the same as the magnitude of $-\mathbf{n}(\mathbf{v},\mathbf{w})$. By definition, $\mathbf{n}(\mathbf{v},\mathbf{w})$ is perpendicular to the plane containing $\mathbf{w}$ and $\mathbf{v}$, and hence so is $-\mathbf{n}(\mathbf{v},\mathbf{w})$. Also, $\mathbf{v}$, $\mathbf{w}$, $\mathbf{n}(\mathbf{v},\mathbf{w})$ form a right-handed system, and so $\mathbf{w}$, $\mathbf{v}$, $\mathbf{n}(\mathbf{v},\mathbf{w})$ form a left-handed system, and hence $\mathbf{w}$, $\mathbf{v}$, $-\mathbf{n}(\mathbf{v},\mathbf{w})$ form a right-handed system. Thus, we have shown that $-\mathbf{n}(\mathbf{v},\mathbf{w})$ is a vector with the same magnitude as $\mathbf{n}(\mathbf{w},\mathbf{v})$ and is perpendicular to the plane containing $\mathbf{w}$ and $\mathbf{v}$, and that $\mathbf{w}$, $\mathbf{v}$, $-\mathbf{n}(\mathbf{v},\mathbf{w})$ form a right-handed system. So by definition this means that $-\mathbf{n}(\mathbf{v},\mathbf{w})$ must be $\mathbf{n}(\mathbf{w},\mathbf{v})$. ✓

Step 5: Show that $\mathbf{n}(\mathbf{v},\mathbf{w}) = \mathbf{v}\times\mathbf{w}$ for all vectors $\mathbf{v}$, $\mathbf{w}$.

Write $\mathbf{v} = v_1\,\mathbf{i} + v_2\mathbf{j} + v_3\,\mathbf{k}$ and $\mathbf{w} = w_1\,\mathbf{i} + w_2\mathbf{j} + w_3\,\mathbf{k}$. Then by Steps 3 and 4, we have

$$
\begin{aligned}
\mathbf{n}(\mathbf{v},\mathbf{w}) &= \mathbf{n}(v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k},w_1\,\mathbf{i}+w_2\,\mathbf{j}+w_3\,\mathbf{k})\\
&= \mathbf{n}(v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k},w_1\,\mathbf{i}) \;+\; \mathbf{n}(v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k},w_2\,\mathbf{j}+w_3\,\mathbf{k})\\
&= \mathbf{n}(v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k},w_1\,\mathbf{i}) \;+\; \mathbf{n}(v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k},w_2\,\mathbf{j}) \;+\; \mathbf{n}(v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k},w_3\,\mathbf{k})\\
&= -\mathbf{n}(w_1\,\mathbf{i},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}) \;+\; -\mathbf{n}(w_2\,\mathbf{j},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}) \;+\; -\mathbf{n}(w_3\,\mathbf{k},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}).
\end{aligned}
$$

We can use Steps 1 and 2 to evaluate the three terms on the right side of the last equation above:

$$
\begin{aligned}
-\mathbf{n}(w_1\,\mathbf{i},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}) &= -\mathbf{n}(w_1\,\mathbf{i},v_1\,\mathbf{i}) \;+\; -\mathbf{n}(w_1\,\mathbf{i},v_2\,\mathbf{j}) \;+\; -\mathbf{n}(w_1\,\mathbf{i},v_3\,\mathbf{k})\\
&= -v_1w_1\mathbf{n}(\mathbf{i},\mathbf{i}) \;+\; -v_2w_1\mathbf{n}(\mathbf{i},\mathbf{j}) \;+\; -v_3w_1\mathbf{n}(\mathbf{i},\mathbf{k})\\
&= -v_1w_1(\mathbf{i}\times\mathbf{i}) \;+\; -v_2w_1(\mathbf{i}\times\mathbf{j}) \;+\; -v_3w_1(\mathbf{i}\times\mathbf{k})\\
&= -v_1w_1\mathbf{0} \;+\; -v_2w_1\,\mathbf{k} \;+\; -v_3w_1(-\mathbf{j})\\
-\mathbf{n}(w_1\,\mathbf{i},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}) &= -v_2w_1\,\mathbf{k} \;+\; v_3w_1\mathbf{j}
\end{aligned}
$$

Similarly, we can calculate

$$
-\mathbf{n}(w_2\mathbf{j},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}) \;=\; v_1w_2\,\mathbf{k} \;-\; v_3w_2\,\mathbf{i}
$$

and

$$
-\mathbf{n}(w_3\mathbf{j},v_1\,\mathbf{i}+v_2\,\mathbf{j}+v_3\,\mathbf{k}) \;=\; -v_1w_3\mathbf{j} \;+\; v_2w_3\,\mathbf{i}\;.
$$

Thus, putting it all together, we have

$$
\begin{aligned}
\mathbf{n}(\mathbf{v},\mathbf{w}) &= -v_2w_1\,\mathbf{k} \;+\; v_3w_1\mathbf{j} \;+\; v_1w_2\,\mathbf{k} \;-\; v_3w_2\,\mathbf{i} \;-\; v_1w_3\mathbf{j} \;+\; v_2w_3\,\mathbf{i}\\
&= (v_2w_3-v_3w_2)\mathbf{i} \;+\; (v_3w_1-v_1w_3)\mathbf{j} \;+\; (v_1w_2-v_2w_1)\mathbf{k}\\
&= \mathbf{v}\times\mathbf{w}\;\; \text{by definition of the cross product.}
\end{aligned}
$$

$\therefore$ $\mathbf{n}(\mathbf{v},\mathbf{w})=\mathbf{v}\times\mathbf{w}$ for all vectors $\mathbf{v}$, $\mathbf{w}$. ✓

So since $\mathbf{v}$, $\mathbf{w}$, $\mathbf{n}(\mathbf{v},\mathbf{w})$ form a right-handed system, then $\mathbf{v}$, $\mathbf{w}$, $\mathbf{v}\times\mathbf{w}$ form a right-handed system, which completes the proof.

Appendix C

3D Graphing with Gnuplot

Gnuplot is a free, open-source software package for producing a variety of graphs. Versions are available for many operating systems. Below is a very brief tutorial on how to use Gnuplot to graph functions of several variables.

INSTALLATION

1. Go to `http://www.gnuplot.info/download.html` and follow the links to download the latest version for your operating system. For Windows, you should get the Zip file with a name such as `gp420win32.zip`, which is version 4.2.0. All the examples we will discuss require at least version 4.2.0.

2. Install the downloaded file. For example, in Windows you would unzip the Zip file you downloaded in Step 1 into some folder (use the "Use folder names" option if extracting with WinZip).

RUNNING GNUPLOT

1. In Windows, run **`wgnuplot.exe`** from the folder (or `bin` folder) where you installed Gnuplot. In Linux, just type **`gnuplot`** in a terminal window.

2. You should now get a Gnuplot terminal with a **`gnuplot>`** command prompt. In Windows this will appear in a new window, while in Linux it will appear in the terminal window where the **`gnuplot`** command was run. For Windows, if the font is unreadable you can change it by right-clicking on the text part of the Gnuplot window and selecting the "Choose Font.." option. For example, the font "Courier", style "Regular", size "12" is usually a good choice (that choice can be saved for future sessions by right-clicking in the Gnuplot window again and selecting the option to update wgnuplot.ini).

3. At the **`gnuplot>`** command prompt you can now run graphing commands, which we will now describe.

GRAPHING FUNCTIONS

The usual way to create 3D graphs in Gnuplot is with the **splot** command:

splot <range> <comma-separated list of functions>

For a function $z = f(x, y)$, <range> is the range of x and y values (and optionally the range of z values) over which to plot. To specify an x range and a y range, use an expression of the form $[a : b][c : d]$, for some numbers $a < b$ and $c < d$. This will cause the graph to be plotted for $a \le x \le b$ and $c \le y \le d$.

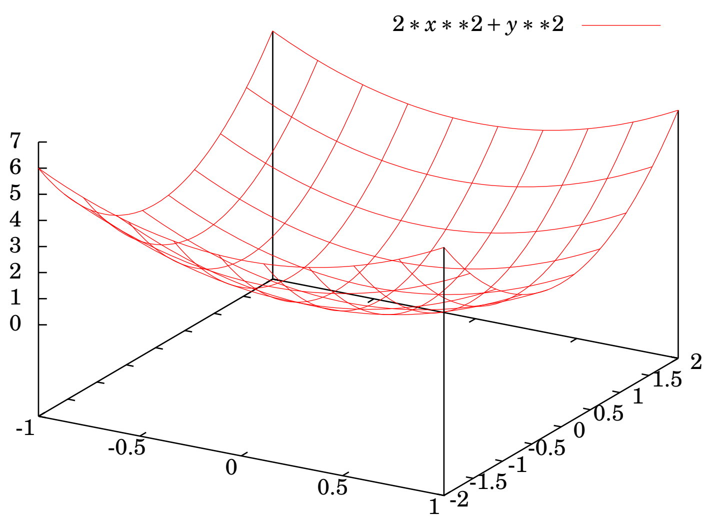
Function definitions use the x and y variables in combination with mathematical operators, listed below:

Symbol	Operation	Example	Result
+	Addition	$2+3$	5
−	Subtraction	$3-2$	1
*	Multiplication	2*3	6
/	Division	4/2	2
**	Power	2**3	$2^3 = 8$
exp(x)	e^x	exp(2)	e^2
log(x)	$\ln x$	log(2)	$\ln 2$
sin(x)	$\sin x$	sin(pi/2)	1
cos(x)	$\cos x$	cos(pi)	-1
tan(x)	$\tan x$	tan(pi/4)	1

Example C.1. To graph the function $z = 2x^2 + y^2$ from $x = -1$ to $x = 1$ and from $y = -2$ to $y = 2$, type this at the **gnuplot>** prompt:

```
splot [-1:1][-2:2] 2*x**2 + y**2
```

The result is shown below:



Note that we had to type 2*x**2 to multiply 2 times x^2. For clarity, parentheses can be used to make sure the operations are being performed in the correct order:

```
splot [-1:1][-2:2] 2*(x**2) + y**2
```

In the above example, to also plot the function $z = e^{x+y}$ on the same graph, put a comma after the first function then append the new function:

```
splot [-1:1][-2:2] 2*(x**2) + y**2, exp(x+y)
```

By default, the x-axis and y-axis are not shown in the graph. To display the axes, use this command *before* the **splot** command:

```
set zeroaxis
```

Also, by default the x- and y-axes are switched from their usual position. To show the axes with the orientation which we have used throughout the text, use this command:

```
set view 60,120,1,1
```

Also, to label the axes, use these commands:

```
set xlabel "x"
set ylabel "y"
set zlabel "z"
```

To show the level curves of the surface $z = f(x, y)$ on both the surface and projected onto the xy-plane, use this command:

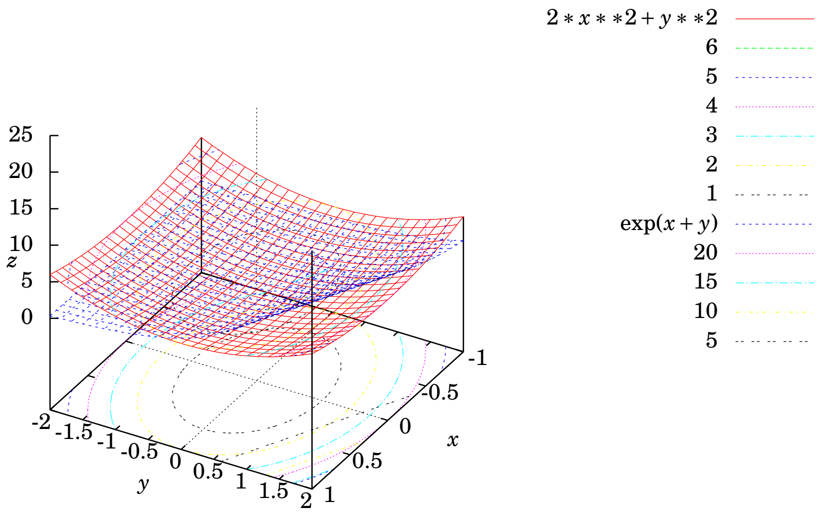
```
set contour both
```

The default mesh size for the grid on the surface is 10 units. To get more of a colored/shaded surface, increase the mesh size (to, say, 25) like this:

```
set isosamples 25
```

Putting all this together, we get the following graph with these commands:

```
set zeroaxis
set view 60,120,1,1
set xlabel "x"
set ylabel "y"
set zlabel "z"
set contour both
set isosamples 25
splot [-1:1][-2:2] 2*(x**2) + y**2, exp(x+y)
```



The numbers listed below the functions in the key in the upper right corner of the graph are the "levels" of the level curves of the corresponding surface. That is, they are the numbers c such that $f(x,y) = c$. Because of the large number of level curves, the key was put outside the graph with the **set key outside** command. If you do not want the function key displayed, it can be turned off with this command: **unset key**

PARAMETRIC FUNCTIONS

Gnuplot has the ability to graph surfaces given in various parametric forms. For example, for a surface parametrized in cylindrical coordinates

$$x = r\cos\theta\ , \quad y = r\sin\theta\ , \quad z = z$$

you would do the following:

```
set mapping cylindrical
set parametric
splot [a:b][c:d] v*cos(u),v*sin(u),f(u,v)
```

where the variable u represents θ, with $a \le u \le b$, the variable v represents r, with $c \le v \le d$, and $z = f(u,v)$ is some function of u and v.

Example C.2. The graph of the helicoid $z = \theta$ in Example 1.34 from Section 1.7 (p. 49) was created using the following commands:

```
set mapping cylindrical
set parametric
set view 60,120,1,1
set xyplane 0
set xlabel "x"
set ylabel "y"
set zlabel "z"
unset key
set isosamples 15
splot [0:4*pi][0:2] v*cos(u),v*sin(u),u
```

The command **set xyplane 0** moves the z-axis so that $z = 0$ aligns with the xy-plane (which is *not* the default in Gnuplot). Looking at the graph, you will see that r varies from 0 to 2, and θ varies from 0 to 4π.

PRINTING AND SAVING

In Windows, to print a graph from Gnuplot right-click on the titlebar of the graph's window, select "Options" and then the "Print.." option. If that does not work on your version of Gnuplot, then go to the File menu on the main Gnuplot menubar, select "Output Device ...", and enter `pdf` in the *Terminal type?* textfield, hit OK. That will allow you to print the graph as a PDF file.

To save a graph, say, as a PNG file, go to the File menu on the main Gnuplot menubar, select "Output Device ...", and enter `png` in the *Terminal type?* textfield, hit OK. Then, in the File menu again, select the "Output ..." option and enter a filename (say, graph.png) in the *Output filename?* textfield, hit OK. Now run your splot command again and you should see a file called graph.png in the current directory (usually the directory where `wgnuplot.exe` is located, though you can change that setting using the "Change Directory ..." option in the File menu).

In Linux, to save the graph as a file called graph.png, you would issue the following commands:

```
set terminal png
set output 'graph.png'
```

and then run your splot command. There are many terminal types (which determine the output format). Run the command **set terminal** to see all the possible types. In Linux, the **postscript** terminal type is popular, since the print quality is high and there are many PostScript viewers available.

To quit Gnuplot, type **quit** at the **`gnuplot>`** command prompt.

GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The **"Document"**, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as **"you"**. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A **"Modified Version"** of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A **"Secondary Section"** is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The **"Invariant Sections"** are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The **"Cover Texts"** are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A **"Transparent"** copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called **"Opaque"**.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The **"Title Page"** means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section **"Entitled XYZ"** means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as **"Acknowledgments"**, **"Dedications"**, **"Endorsements"**, or **"History"**.) To **"Preserve the**

Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgments" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgments and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the

combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgments", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of

some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgments", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright ©YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

History

This section contains the revision history of the book. For persons making modifications to the book, please record the pertinent information here, following the format in the first item below.

1. VERSION: 1.0

   Date: 2008-01-04

   Author(s): Michael Corral

   Title: Vector Calculus

   Modification(s): Initial version

Index

Symbols

A

B

C

J

L

M

N

O

P

Q

R

S

T

U

V

W

Z